Facial Analysis: Looking at Biometric Recognition and Genome-Wide Association

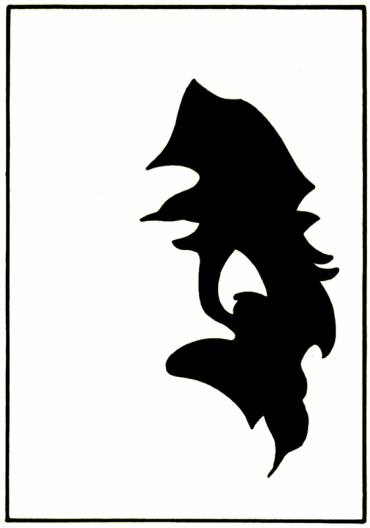
Fagertun, Jens

Publication date:
2013

Document Version
Publisher's PDF, also known as Version of record

Facial Analysis: Looking at Biometric Recognition and Genome-Wide Association

Jens Fagertun



Kongens Lyngby 2013
PhD-2013-319

Technical University of Denmark
Department of Applied Mathematics and Computer Science
Building 321, DK-2800 Kongens Lyngby, Denmark
Phone +45 45253351, Fax +45 45882673
reception@compute.dtu.dk
www.compute.dtu.dk PhD-2013-319

Summary

The goal of this Ph.D. project is to present selected challenges regarding facial analysis within the fields of *Human Biometrics* and *Human Genetics*. In the course of the Ph.D. nine papers have been produced, eight of which have been included in this thesis.

Three of the papers focus on face and gender recognition, where in the gender recognition papers the process of human perception of gender is analyzed and used to improve machine learning algorithms.

One paper addresses the issues of variability in human annotation of facial landmarks, which most papers regard as a static "gold standard". However, we document intra- and inter-operator variability associated with annotating these landmarks, which is a valuable result for applications that are sensitive to such variability.

One paper presents a comprehensive proof-of-concept study of the prediction of facial characteristics based solely on genetic information, a new area that holds great potential.

Two papers explore the connection between minor physical anomalies in the face and schizophrenic disorders. Schizophrenia is a life long disease, but early discovery and treatment can have a significant impact on the course of the disease.

Finally, one paper presents a novel appearance model that is a fusion of the active appearance models and the Riemannian elasticity framework.

Resumé

Målet for denne afhandling er at præsentere udvalgte problemstillinger vedrørende ansigtsanalyse inden for områderne *Menneskelig Biometri* og *Menneskelig Genetik*. Der er i løbet af dette ph.d.-studie blevet udarbejdet ni artikler, hvoraf otte er inkluderet i denne afhandling.

De første tre artikler fokuserer på ansigts- og kønsgenkendelse. I artiklerne omhandlende kønsgenkendelse analyseres det, hvordan mennesket opfatter og analyserer køn, hvilket efterfølgende bruges til at forbedre maskinlæringsalgoritmerne.

Den fjerde artikel diskuterer spørgsmålet om den variation, der er i menneskelige annotation af ansigtslandmærker, som de fleste artikler betragter som en statisk "guldstandard". I artiklen belyser vi den interne og externe operatørvariation, der er forbundet med annotering af disse landmærker. Dette resultat er værdifuldt for de applikationer, som er følsomme over for denne variation.

Den femte artikel præsenterer et omfattende "proof-of-concept"-studie omkring forudsigelsen af ansigtstræk udelukkende baseret på genetisk information. Dette er et nyt område, der rummer et stort potentiale.

I de to følgende artikler udforskes det, om der er en sammenhæng mellem mindre fysiske abnormiteter i ansigtet og skizofrene lidelser. Skizofreni er en livslang sygdom, men tidlig opdagelse og behandling kan have en markant indflydelse på sygdomsforløbet.

I den sidste artikel præsenteres en ny "appearance model", som er en sammensmeltning af de to metoder "active appearance models" og det "Riemannian

elasticity framework".

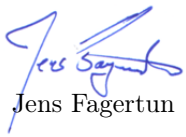
Preface

This thesis was prepared at the Image Analysis and Computer Graphics section at the Department of Applied Mathematics and Computer Science at the Technical University of Denmark (DTU). It was done in fulfilment of the requirements for acquiring a doctor of philosophy degree (Ph.D.) within the topic of image analysis. The project was funded by DTU.

The research presented in this thesis deals with facial analysis in two fields: *Human Biometrics* and *Human Genetics*. The thesis consists of introductory parts to both fields, overview of the applied methods, a technical description of the software packages and hardware setups created for data collection and a concluding part. Following the conclusion are the eight manuscripts prepared during the course of the Ph.D. study.

The project has been supervised by Associate Professor Rasmus R. Paulsen at DTU. The research has mainly been carried out at DTU but also at the division of Central Nervous System diseases at deCODE Genetics, Iceland.

Lyngby, 1st of December-2013



Jens Fagertun

Papers Included in the Thesis

[A] J. Fagertun, D.D. Gomez, M.F. Hansen, and R.R. Paulsen. Sparse similarity-based fisherfaces. In A. Heyden and F. Kahl, editors, *Image Analysis*, volume 6688 of *Lecture Notes in Computer Science*, pages 69–78. Springer Berlin Heidelberg, 2011

[B] J. Fagertun, T. Andersen, and R.R. Paulsen. Gender recognition using cognitive modeling. In A. Fusiello, V. Murino, and R. Cucchiara, editors, *Computer Vision – ECCV 2012. Workshops and Demonstrations*, volume 7584 of *Lecture Notes in Computer Science*, pages 300–308. Springer Berlin Heidelberg, 2012

[C] J. Fagertun, T. Andersen, T.F. Hansen, and R.R. Paulsen. 3D gender recognition using cognitive modeling. In *2013 International Workshop on Biometrics and Forensics (IWBF)*. IEEE, 2013

[D] J. Fagertun, S. Harder, A. Rosengren, C.K. Møller, T. Werge, R.R. Paulsen, and T.F. Hansen. Variability of 3D facial landmarks. *submitted to Medical Image Analysis*, 2013

[E] J. Fagertun, T. Pers, K. Wolffhechel, R.R. Paulsen, and H. Jarmer. Correlation of complex polygenic variations to specific holistic facial characteristics. *submitted to PLoS ONE*, 2014

[F] T.F. Hansen, J. Fagertun, L. Duong, A. Rosengren, E. Sørensen, H. Ullum, R. Larsen, R.R. Paulsen, and T. Werge. Discrete facial feature of NRNX1 deletions carriers. In *submission*

[G] J. Fagertun, C.K. Møller, T. Werge, R.R. Paulsen, and T.F. Hansen. Detection of minor physical anomalies in the face associated with severe schizophrenic disorders. In *submission*

[H] M.F. Hansen, J. Fagertun, and L. Larsen. Elastic appearance models. In *Proceedings of the British Machine Vision Conference*. BMVA Press, 2011

Papers not Included in the Thesis

[164] K. Wolffhechel, J. Fagertun, U.P. Jacobsen, W. Majewski, A.S. Hemmingsen, C.L. Larsen, S.K. Lorentzen, and H. Jarmer. Interpretation of appearance: The effect of facial features on first impressions. *submitted to PLoS ONE*, 2014

Supervised Projects and Theses

[115] C.K. Møller. 3-Dimentionel craniofacial image analysis: Detection of anthropometric traits as a marker for genetically caused disease risk. *B.Sc.Eng. Thesis*, Copenhagen University Hospital, 2013

[61] M. Frisenfeldt. 3D reconstruction of the human face. *B.Sc.Eng. Thesis*, Technical University of Denmark, 2013

[163] K. Wolffhechel. Can you judge a book by its cover? - an investigation into the proposed connection between looks and personality. *M.Sc.Eng. Thesis*, Technical University of Denmark, 2012

[138] A. Søgaard and S. Bohøj. 3D feature extraction of 3D human facials scans. *B.Sc.Eng. Thesis*, Technical University of Denmark, 2012

[39] J. Dahl. Analysis pipeline for face recognition. *B.Sc.Eng. Thesis*, Technical University of Denmark, 2012

[105] W. Majewski. Identification of the optimal method for analysis of human faces in the context of finding associations to personality traits. *Special Course*, Technical University of Denmark, 2012

[132] S. Sahin and J.H. Saabye. The connection between 3D facial scans and schizophrenia. *Special Course*, Technical University of Denmark, 2011

Acknowledgements

During my three years as a Ph.D. student I have had the honor to collaborate, work, discuss, research and socialize with dedicated and inspiring people. I would like to thank the following people for their support and invaluable assistance in the preparation of the work presented in this thesis:

First and foremost, I thank my supervisor Rasmus R. Paulsen for his support throughout this thesis.

I would like to thank my Ph.D. assessment committee associate professor Line K. H. Clemmensen, professor Timothy Cootes and associate professor Dan Witzner Hansen, for their thorough review and constructive feedback.

I would also like to thank all my collaborators throughout this project: Thomas F. Hansen and Thomas Werge and the rest of the Institute for Biological Psychiatry at the Copenhagen University Hospital, Denmark for their dedication to our collaboration; Hreinn Stefansson and the Central Nervous System diseases division at deCODE Genetics, Iceland for the time they all spent working with me; Hanne Ø. Jarmer and the Behavioral Phenomics group at Center for Biological Sequence Analysis at the Technical University of Denmark, Denmark for our collaboration; Tobias Andersen from the Cognitive Systems section at the Department of Applied Mathematics and Computer Science at the Technical University of Denmark, Denmark for our collaboration; David D. Gomez from the Department of Signal Theory and Communications at Carlos III University, Spain for our work on face recognition; Mads F. Hansen for our work together and spending considerable time on the AAMLab software package.

Further on, I would like to thank Allan A. Nielsen, Line K. H. Clemmensen, Rasmus Larsen and Knut Conradsen for their great patience when answering questions of statistical nature. Also thanks to J. Andreas Bærenten for all the discussions on mesh processing.

Thanks to my colleagues for using time to participate in the cognitive experiments: Asger N. Christiansen, Christian T. Larsen, Lasse F. Laursen and Stine Harder.

A special thanks to Otto H. A. Nielsen, Jannik B. Nielsen, Anders B. L. Larsen, Ulla Jensen and Anna M. Fagertun for reviewing various sections of this thesis.

Finally, I would like to thank all my past and present colleagues at the Image Analysis and Computer Graphics Section at the Department of Applied Mathematics and Computer Science at the Technical University of Denmark for providing a pleasant and inspiring atmosphere, where there is always time to answer and discuss questions.

In short it has been an extraordinary experience and great opportunity to work with all of you.

Contents

Chapter 1

Introduction

Face analysis, in the form of face detection and face recognition, is a task so innate to humans, that individuals do not even notice the extensive number of times it is performed every day. From a very early stage the human brain is capable of detecting facial features from limited amount of visual information. Craig Mooney, a cognitive psychologist, discovered this in 1957 [116] by binary facial images with very little visual information currently known as "Mooney faces". Such images are usually very hard to perceive as faces if viewed upside down (see Figure 1.1).

Figure 1.1: Examples of "Mooney faces" turned upside down. Courtesy of `pics.stir.ac.uk`.

However, this ability to detect and extrapolate facial features from very little amounts of visual clues also makes us see faces "everywhere". A collection of this phenomenon can be seen in Figure 1.2.

Figure 1.2: Phenomena in nature perceived as faces. Courtesy of `Bill Snell, ancient-wisdom.co.uk, Ranger Rick magazine, NASA - The Viking Project and Trond Blomlie.`

Even though humans are experts in facial perception it is not yet fully understood how this analysis is performed. For many years psychophysicists and neuroscientists have been researching whether face recognition is done holistically or by local feature analysis, i.e., is face recognition done by looking at the face as a whole or by looking at local facial features independently [15, 64]. It is however clear that humans are only capable of holding one face image in the mind at a given time. Figure 1.3 shows a classical illusion called "The Wife and the Mother-in-Law", which was introduced into the psychological literature by Edwin G. Boring. What do you see? A witch or a young lady?

In this thesis we work with facial analysis within the fields of *Human Biometrics* and *Human Genetics*. The thesis starts with examining the human face perception capabilities and use this in the field of *Human Biometrics* and then it delves into the field of *Human Genetics* by researching the link between the human genome and facial morphology. It should be mentioned that having some knowledge in the fields of image analysis and genetics will make this work more accessible for the reader. This work is highly interdisciplinary and it draws on theory from multivariate statistics, numerical analysis, linear algebra, genetics, computational geometry, computer science, etc. By covering such a broad spectrum of topics this thesis will inevitably lack depth in some of the covered

Figure 1.3: "The Wife and the Mother-in-Law" by Edwin G. Boring. What do you see? A witch or a young lady? Courtesy of Danial Chandler.

domains. As a result, the treatment of particular topics may appear insufficient to some. The author has sought to redeem this by providing pointers to relevant literature sources.

1.1 Objectives

Because face analysis is such an intrinsic part of human interaction, it has been studied in many fields from different perspectives. In computer science face analysis has been conducted since the 1960's and many face analysis and face modeling techniques have been devised during this time, such as face detection [157], active appearance models [33], eigen- [151], Fisher- [12] and Laplacian-faces [79], etc. Face analysis is a big part of *Human Biometrics* and lately it has begun to spread into the field of *Human Genetics* research. This thesis presents novel work in facial analysis within the fields of *Human Biometrics* and *Human Genetics*, trying to answer a few of the many missing links in these fields.

The objectives of this thesis are to:

- Progress our understanding of biometric recognition based on facial data

from a human point of view.

- Progress our understanding of the human genome with respect to facial morphology.
- Assess the potential of detecting severe schizophrenic disorders from minor physical anomalies in the face.
- Contribute to the construction of data sets containing both high resolution 3D scans and genotype data.

1.2 Thesis Overview

This thesis consists of three main parts: a part about biometric recognition, a part about associations between the human genome and the face, and finally a part about the software packages and hardware implementations constructed as part of this work. Eight research papers have been produced and are found in appendix. These papers, together with the work described in the implementation part, represent the major contribution of this thesis. The first two parts serve as introduction and give historical perspective in order to motivate the choices made in the papers. The papers are self-contained and can be read independently. The rest of the thesis should be read in succession.

Part I Biometrics presents an introduction and historical perspectives to the field of biometric recognition. It discusses the general biometric recognition process and the different considerations regarding biometric recognition. Furthermore, it presents core methods used in the contributions related to *Human Biometrics*.

Part II Genetics presents an introduction to *Human Genetics* and to *Heritability of Schizophrenia* for the unacquainted reader. It presents core methods used in the contributions related to *Human Genetics*, such as genome-wide association studies.

Part III Implementation documents the design and development of several software packages and tools together with the design of hardware setups for collecting 2D images and 3D facial scans.

Part IV Discussion and Conclusion presents a discussion of the work done in this thesis and ends with concluding remarks.

Part V Appendices

Appendix A explores an extension to the similarity based fisherfaces algorithm with sparse Principal Component Analysis.

Appendix B introduces how human cognition can improve machine learning algorithms on low resolution texture data in a gender recognition setup.

Appendix C describes how human cognition can improve machine learning algorithms on high resolution shape and texture data in a gender recognition setup.

Appendix D describes the variability of human annotated facial landmarks.

Appendix E introduces a new framework for understanding and predicting facial characteristics based solely on genetic information.

Appendix F explores discrete facial feature of NRNX1 deletions carriers.

Appendix G identifies associations between minor physical anomalies in the face and severe schizophrenic disorders.

Appendix H presents a fusion of the active appearance model and the Riemannian elasticity framework, which yields a non-linear shape model and a linear texture model.

1.3 Mathematical Notation

Throughout this thesis the following mathematical notations are used:

Scalar values are denoted with lower-case Latin or Greek letters:

$$x$$

Vectors are denoted with lower-case, bold Latin or Greek letters. In this thesis only column vectors are used:

$$\mathbf{x} = [x_1,\ x_2,\ \ldots,\ x_n]^T$$

Matrices are denoted with capital, non-italic bold Latin or Greek letters:

$$\mathbf{X} = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$$

The sample mean vector of a specific data set is denoted with lower-case, bold Latin or Greek letters with a bar:

$$\bar{\mathbf{x}}$$

1.4 Abbreviations

A list of the abbreviations used in this thesis can be found below:

AA	Archetype Analysis.
AAM	Active Appearance Models.
ASM	Active Shape Model.
BMI	Body Mass Index.
CIR	Correct Identification Rate.
CNV	Copy-Number Variation.
DBDS	Danish Blood Donor Study.
DNA	Deoxyribonucleic Acid.
EAM	Active Elastic Appearance Model.
FAR	False Acceptance Rate.
FIR	False Identification Rate.
FRR	False Rejection Rate.
GSR	Gender Strength Regression.
GWAS	Genome-Wide Association Study.
ICA	Independent Component Analysis.
PCA	Principal Component Analysis.
LDA	Linear Discriminant Analysis.
MAF	Minimum / Maximum Autocorrelation Factor Analysis. Minor Allele Frequency.
MDL	Minimum Description Length.
MPA	Minor Physical Anomalies.
MRF	Markov Random Field.
NMF	Non-Negative Matrix Factorization.
PA	Procrustes Analysis.
PDM	Point Distribution Model.
SD	Standard Deviation.
SNP	Single-Nucleotide Polymorphism.
SPC	Sparse Principal Component.
SPCA	Sparse Principal Component Analysis.
SVM	Support Vector Machine.

Chapter 2

Contributions

As part of this thesis eight contributions have been produced: one contribution focuses on face recognition, two contributions focus on gender recognition, one contribution on the variability of 3D facial landmarks, one contribution with the ambitious goal of making a facial portrait from DNA, two contributions about facial features associated with severe schizophrenic disorders and finally one contribution describing the Elastic Appearance Models. They are introduced in the following sections and can be found in full length in Appendix A, B, C, D, E, F, G and H.

2.1 Sparse Similarity-based Fisherfaces

The author's previous work [53, 68] was extended in *Sparse Similarity-based Fisherfaces* [A]. Sparse Similarity-based Fisherfaces is a novel facial recognition algorithm that aims at mimicking the human ability to discriminate faces by projecting the faces in a highly discriminative and easy interpretative way. Pixel intensities are used by Sparse Principal Component Analysis and Linear Discriminant Analysis to assign a one dimensional subspace projection to each person belonging to a reference data set. Experimental results performed in the public AR data set show that Similarity-based Fisherfaces in a sparse version can obtain the same recognition results as the technique in a dense version using

only a fraction of the input data.

Two distinct aspects from the contribution merit particular note:

Interpretation vs. performance. The idea of using SPCA in facial recognition is based on the construction of a basis of sparse facial components that are more human understandable than the global structure of PCA. Here the finding suggests that performance-wise PCA exhibits high robustness, requires no tuning of parameters and there appears to be little or no gain from using a more complex method. Interpretation-wise SPCA offers a more understandable basis with local changes as opposed to the global changes of PCA. Thus, the dimension reduction method should be chosen depending on the application and whether performance or interpretation is desired. Similar results are obtained when investigating other methods such as NMF, AA and ICA. MAF appears to perform very similar to PCA.

Robustness to occlusions. In the contribution it is speculated whether or not SPCA compared to PCA offers robustness to occlusions. This has been investigated later by a student in [39] supervised by the author. The project concluded that it does not appear to be the case.

2.2 Gender Recognition using Cognitive Modeling

In *Gender recognition using cognitive modeling* [B] we suggest to view the biometric gender recognition problem from a cognitive modeling point of view. We estimate a *"gender strength"* of frontal faces, which is a continuous class variable, superseding the traditional binary class labeling. To incorporate this continuous variable we suggest a novel linear gender classification algorithm, the Gender Strength Regression. In addition, we use the gender strength to construct a smaller but refined training set, by identifying and removing ill-defined training examples. We use this refined training set to improve the performance of known classification algorithms. The human performance of gender recognition on known data sets is reported, and surprisingly it appears to be a difficult task for humans. Finally our results are reproduced on a data set of more than 40,000 public Danish LinkedIN profile pictures.

The following aspects from the contribution merit particular note:

Human cognitive performance. In time constrained test, where the test

participants only were exposed to an image for 200ms, the performance was about ~ 20 per cent misclassified. The test participants reported that they felt as if they were guessing most of the time, but this is clearly better than guessing, which is a remarkable example of the human cognition capability.

Training set quality. The results show that not only the volume of the training data is important but also the quality of the training data. This important and interesting issue is not yet discussed very much in the literature. The presented method for identifying ill-defined training samples is time consuming and probably not optimal. How to optimally determine ill-defined training samples and more interestingly how to identify them without an intensive user interaction is a topic open for further research.

Generalization performance. In order to see how the algorithms generalize to unseen and unconstrained data, a huge data set of 40,692 images from public Danish LinkedIN profiles was used. The results showed a drastic reduction in false classifications when employing the training set without ill-defined samples.

2.3 3D Gender Recognition using Cognitive Modeling

Contribution [B] was extended in *3D Gender recognition using cognitive modeling* [C]. Here we use 3D scans of human faces and compared to [B] we now look at both the shape and the texture[1]. Furthermore, the scans were in higher resolution than the images in [B]. As in [B] we estimate the *"gender strength"*. That enables us to visualize some of the visual trends humans use when performing gender classification and to construct a smaller but refined training set. Also as in [B] we use this refined training set to improve the performance of known classification algorithms.

The following aspects from the contribution merit particular note:

Discriminating power of shape and texture. When looking at the shape and texture experiments in the contribution it is clear that texture holds much more discriminating power than shape in gender recognition.

[1]The texture values are sampled the same place between individuals by means of point correspondence.

Humans are outperformed. Interestingly, with high resolution 3D facial scans, standard machine learning algorithms such as LDA and SVM outperform humans on data acquired in constrained environment. It remains to be seen if this remains valid also in more unconstrained environments. This is opposed to the results in [B] where human and machine performances were similar.

Difference in human classification of male and female. Interestingly, both in this contribution and in [B] there appears to be a clear difference if male or female are wrongly classified. Females were more frequently classified wrongly as males. The reason for this is not investigated in this work.

2.4 Variability of 3D Facial Landmarks

In *Variability of 3D facial landmarks* [D] a full facial 3D annotation procedure is presented for a sparse set of manually annotated landmarks. Furthermore, an extended set of manually annotated landmarks are analyzed with focus on intra- and inter-operator variability. We then select a sparse set of landmarks in order to construct dense correspondence of the 3D scans with a minimum of point variance. Using the sparse set of landmarks, we minimize the point mean variance. Finally, we provide a 3D full facial map as a reference for annotation variation of other putative landmarks using this sparse landmark set.

The following aspects from the contribution merit particular note:

Inter- and intra-operator variability. In this study we analyze the inter- and intra-operator variability of facial 3D landmarks.

Quality of dense point correspondence. We evaluate and visualize the quality of dense point correspondence from different sparse landmark selections.

Generalizability limitation. We used subjects that are Caucasian with Scandinavian background, thus we cannot exclude that the variability of the annotation landmarks is different for other ethnicities, e.g., the texture of blonde eyebrows on light skin may be difficult to separate, whereas dark eyebrows may not.

2.5 Correlation of Complex Polygenic Variations to Specific Holistic Facial Characteristics

In *Correlation of complex polygenic variations to specific holistic facial characteristics* [E] a proof-of-concept study of the prediction of facial characteristics based solely on genetic information is presented. Here for the first time we attempt a holistic approach, where we correlate complex polygenic variations to specific holistic facial characteristics across a population of 1,338 individuals by finding a one-way mapping between a genome-wide statistical SNP model and a facial statistical shape model. We show that it is possible to predict several components in the facial statistical shape model from genetic data, where we get a statistically significant correlation (Bonferroni corrected) between actual and predicted values. For the best performing facial components we are able to classify to simple facial archetypes better than random.

Two distinct aspects from the contribution merit particular note:

Proof-of-concept study. To our knowledge, this is the first study to show a correlation of complex polygenic variations to specific holistic facial characteristics.

Sample size. The signal strength is such that we have a moderate Pearson correlation and perform better than random in classification experiments (in the best cases). It is hard to conclude anything other than that we see a signal in this proof-of-concept study, but it is unclear how strong this signal can be with a larger cohort above the $\sim 1,300$ individuals we currently are working with.

2.6 Discrete Facial Feature of NRNX1 Deletions Carriers

In *Discrete facial feature of NRNX1 deletions carriers* [F] it is examined whether there are specific facial traits for carriers of copy number variations in the neurexin 1 gene. We focused on facial features, as previous studies have found a general increase in minor physical anomalies in the face among patients with schizophrenia and autism. Images from two unrelated families with members affected by severe mental disorders with six carriers of deleterious mutations in neurexin 1, and 605 unrelated healthy blood donors are analyzed using a Mann-Whitney test. We found a nominal significant thicker upper lip, vermillion, and

a smaller nose width of individuals carrying a deleterious mutation in neurexin 1 as compared with their relatives who do not carry deleterious mutations in neurexin 1.

The following aspects from the contribution merit particular note:

Explorative study - novelty. To our knowledge, this is the first study to show a discrete facial trait to be associated with deleterious NRXN1 mutations.

Explorative study - small sample size. The study comprises of only a small sample size and does not include replication of the findings in an independent sample, which makes this study have insufficient statistical power. Nevertheless, there are many indications for a true finding. For example, the association was present when analyzing two independent families and different generations, and the finding does not associate between family-membership. Furthermore, the two families are quite unique and collection of larger samples is very hard to accomplish.

2.7 Detection of Minor Physical Anomalies in the Face Associated with Severe Schizophrenic Disorders

In *Detection of minor physical anomalies in the face associated with severe schizophrenic disorders* [G] an objective analysis of the connection between minor physical anomalies in the face and mental disorders within the schizophrenic spectrum is presented. We use high resolution 3D scans to associate human understandable distances to the disorder. In a classification experiment we show very good performance in distinguishing patients from controls. Finally, we construct a "face map" that shows regions that associate with the schizophrenic disorder spectrum.

The following aspects from the contribution merit particular note:

Association and performance. The experiments conducted in this work show a strong association between control/patient status and several facial distances, the mouth being the most predominant with a p-value of $3.52 \cdot 10^{-35}$. The classification performance is shown to be as high as 87.4% when separating the controls from the patients.

Extremal control sampling. The control cohort used in this study consists of blood donors. Blood donors are more healthy than the average population, thus reported results will probably be lower if exposed to data that include samples from the general public.

Unknown patient sampling. The schizophrenic spectrum [1] is a syndrome, which consists of a spectra of phenotypes and we cannot be sure if the cohort we work with is representative. Moreover, the different diagnose categories in the schizophrenic spectrum change over time and as diagnosing schizophrenia is a subjective process there can be subjective biases in the diagnoses. All this indicates that our approach to look at the total schizophrenic spectrum is the correct way to perform analysis in the patient/control setup.

Association face map. By using dense point correspondence we build a "face map" which illustrates where on the face the minor physical anomalies associated with schizophrenic disorders are likely to be located.

2.8 Elastic Appearance Models

In *Elastic Appearance Models* [H] we present a fusion of the active appearance model (AAM) and the Riemannian elasticity framework, which yields a non-linear shape model and a linear texture model – the active elastic appearance model (EAM). The non-linear elasticity shape model is more flexible than the usual linear subspace model, and it is therefore able to capture more complex shape variations. This additional flexibility is due mostly to local rotation and local translation invariance characteristics inherent to the model. In addition, we introduce global scale invariance into the Riemannian elasticity framework, which together with the local translation and local rotation invariance eliminate the need for separate pose estimation. The new approach was tested against AAM in three experiments: face labeling, face labeling with poor initialization and corpus callosum segmentation. In all of the examples the EAM performed significantly better than the AAM.

The following aspects from the contribution merit particular note:

Accuracy vs. efficiency. When assessing the differences between EAM and AAM it can be argued that we have gained robustness and accuracy but have lost efficiency compared to the AAM method.

Ground truth. The evaluation done in this work relies on using human annotation as ground truth. As shown in [D] there can be a large inter/intra-

operator variance in human annotations. Nevertheless, the potential bias of the ground truth towards AAM or EAM was not investigated in this work.

Matlab implementation. As part of this contribution we have provided an AAM/EAM Matlab implementation that can be forwarded on request.

Chapter 3

Contributions in Preparation

This chapter presents the contributions of the genetic part of this thesis, which are not yet ready or are not suitable for publication at this time. In the following, references are made to two data sets: the Danish cohort and the Icelandic cohort. The Danish cohort consists of about 1,000 healthy and unrelated subjects recruited among unpaid volunteers of the Danish Blood Donor Study [122]. The setup for recording the image data is described in Section 8.5. The Icelandic cohort consists of about 1,300 individuals recruited at deCODE Genetics and the setup for recording the image data is described in Section 8.3. The genetic analysis (genotyping) is done by chip-typing with Illumina chips at deCODE Genetics for both cohorts.

3.1 Facial Morphology Associated with PAX3, TP63 and C5orf50

Together with the Institute for Biological Psychiatry[1] and deCODE Genetics, the author performed a replication study on two GWA studies of Facial Morphol-

[1]Mental Health Center Sct Hans, Copenhagen University Hospital

ogy. Paternoster *et al.* [119] have reported a correlation between SNPs in the TP63 and C5orf50 genes and two facial distances: "Distance between eyes" and "Left eye to nasion". Liu *et al.* [102] have reported a correlation between an SNP in the PAX3 gene and the facial distance: "Nasion to mid-endocanthion". The "mid-endocanthion" is the middle point between the left- and right-endocanthion (denoted left- and right-medial canthus in this work). A graphical representation of the points is shown in Figure 3.1. A summary of the results from [119] and [102] that were used for replication is given in Table 3.1.

Article	Gene	SNP	Phenotype	n	p-value
[119]	TP63	rs17447439	Distance between eyes	~ 3800	0.043^*
[119]	C5orf50	rs6555969	Left eye to nasion	~ 3800	0.031^*
[102]	PAX3	rs7559271	Nasion to mid-endocanthion	~ 1600	$4 \cdot 10^{-7}$

* Bonferroni corrected p-value.

Table 3.1: Summary of the results from [119] and [102] used for replication.

The replication used SNP and frontal facial images from the Icelandic cohort. 32 participants were removed due to non-neutral facial expression, or non-frontal face direction to the camera, resulting in a remainder of 1,306 participants. For a replication there is no need to correct for multiple testing, so the statistical significant threshold was set to 0.05.

3.1.1 Discussion

Unexpectedly, none of the phenotypes (distances) replicated, even though the original signals reported in [102, 119] appeared strong and the studies used a two-step approach with good power. The causes of non-replication can be many. Two possibilities are the smaller sample size and/or possible noise in the face distances. These should be addressed before concluding non-association. Hence, at this point we are not writing a non-replication article.

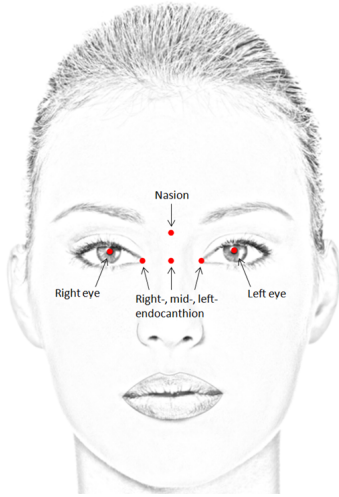


Figure 3.1: A graphical representations of the points used for replication.

3.2 Facial Morphology of CNV Carriers

Together with deCODE Genetics, facial morphology analysis was performed on CNV carriers (associated with schizophrenia) and controls in the Icelandic cohort. This analysis included performing T-tests on the human understandable facial distances described in Section 7.4 and the CNV carriers status. The strongest signal was found between the CNV 15q11.22 duplication[2] and the width of the face (H10) with a p-value of 0.0005. Even though this is a statistically significant result, we determined that the signal strength was not sufficient for a publication at this point. As deCODE are continually gathering more samples we hope in the future to be able to conduct the analysis with a resulting stronger signal. We are currently also working on the Danish cohort where we have identified several CNV carriers in the cohort.

[2]Found to be associated with schizophrenia [143].

3.3 GWAS on Facial Morphology

Together with the Institute for Biological Psychiatry and deCODE Genetics a full GWAS was performed on the human understandable facial distances described in Section 7.4 and about twenty million SNPs on the Icelandic cohort. Here a correlation between Nasion point and upper lip (V19) and a cluster of 18 SNPs in the homo sapiens ceramide synthase genes was found to be genome-wide significant (corrected for multiple testing), the strongest having a p-value of $1.59 \cdot 10^{-11}$. In the deCODE setup, a p-value below 10^{-8} is considered genome-wide significant. A Manhattan plot over the 174,000 most correlated SNPs over all the genome is shown in Figure 3.2, which shows this very promising result. However, a replication study on a new cohort has to be positive before deCODE will release the finding for publication. This is due to the fact that GWAS are prone to have low statistical power when smaller cohorts are used, and therefore vulnerable to the "winner's curse" that describes that the reported association in smaller discovery studies can be exaggerated relative to the association found in follow-up studies. To counter this problem either much more data needs to be collected or a replication study needs to be performed.

A replication attempt has been made with the Institute for Biological Psychiatry on the Danish cohort. The replication cohort (the Danish cohort) does not have a similar setup as the original cohort (the Icelandic cohort). The original cohort consists of 2D images and the Danish cohort consists of 3D scans. Unfortunately the 3D did not replicate the 2D findings (p-value > 0.4). We are currently on the lookout for a 2D cohort recorded under similar conditions to make a replication attempt. The status at the time of writing is that this finding cannot be published and thus, the SNP positions will not be reported here.

3.4 Future Work

In this work substantial amount of data has been accumulated that could open interesting research potential in future works. The following is a list of research topics currently being explored together with the Institute for Biological Psychiatry:

- *Twin studies with schizophrenia*
  We have collected 14 twins, both monozygotic and dizygotic. Both concor-

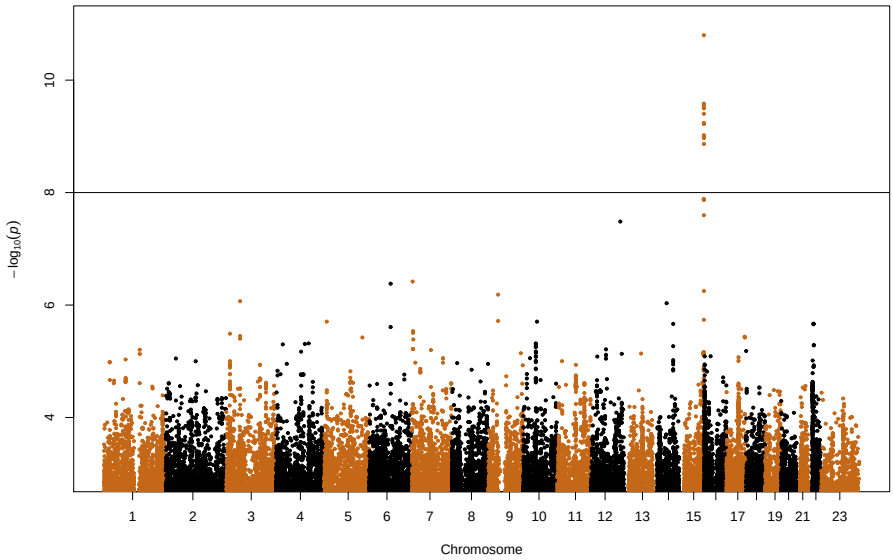


Figure 3.2: Manhattan plot over the 174,000 most correlated SNP's and facial distance V19. Points above the horizontal line indicate genome-wide significance (10^{-8}). Note: chromosome 23 is the X chromosome.

dant[3] and the interesting discordant[4] cases are present among the twins. It is important to note that even though monozygotic twins have a very similar genome it is not exactly the same [104]. Research into these twins both from a facial morphology and from genetic aspects is very interesting.

- *CNV carriers status association to facial morphology*
  Initial analysis of the Danish cohort with respect to CNVs associated with increased risk of developing psychiatric disorders, such as: 1q21.1, 2p16, 3q29, 7q36.3, 15q11.2, 15q13.3, 16p13.11, 17q12, 22q11.2, 16p11.2 and 1q42.2 has been completed. It has shown at least seven of these CNV are present in the Danish cohort.

- *Genetic impact on facial morphology*
  We are continuing the research presented in this work on how genetic variation impacts facial morphology on the Danish cohort with high resolution 3D scans. At the time of writing additional samples are still being collected with the software/hardware setups developed as part of this thesis. Expanding the Danish cohort is of paramount importance to gain the statistical power needed to find associations in future research. Furthermore, we will shortly start to collect samples from children with high risk of developing schizophrenia.

[3] Both twins have schizophrenia.
[4] Only one of the twins have schizophrenia.

Part I

Facial Analysis in a Biometric Context

Chapter 4

Biometrics

When using the word "biometrics" today, it is commonly understood as "human" biometrics. Historically, biometrics have been used within the field of biostatistics, a portmanteau of biology and statistics. The term "biometrics" is derived from the two Greek words "bio" that translates into 'life' and "metric" that translates into 'to measure'. In the remaining of this work we will adopt the term biometrics to refer to human biometrics or biometric authentication as Jain *et al.* [88] suggest it to be called.

4.1 History of Biometrics

The earliest recordings of biometrics are from around 29,000 BC when cavemen would sign their drawings with handprints, see Figure 4.1. From these handprints researchers have shown that both male and female artists painted these images. In 500 BC Babylonian business transactions were signed in clay tablets with fingerprints. In China about 200 BC thumb prints were also found on clay seals.

The first documented biometric identification is by Henry Faulds, a British sur-

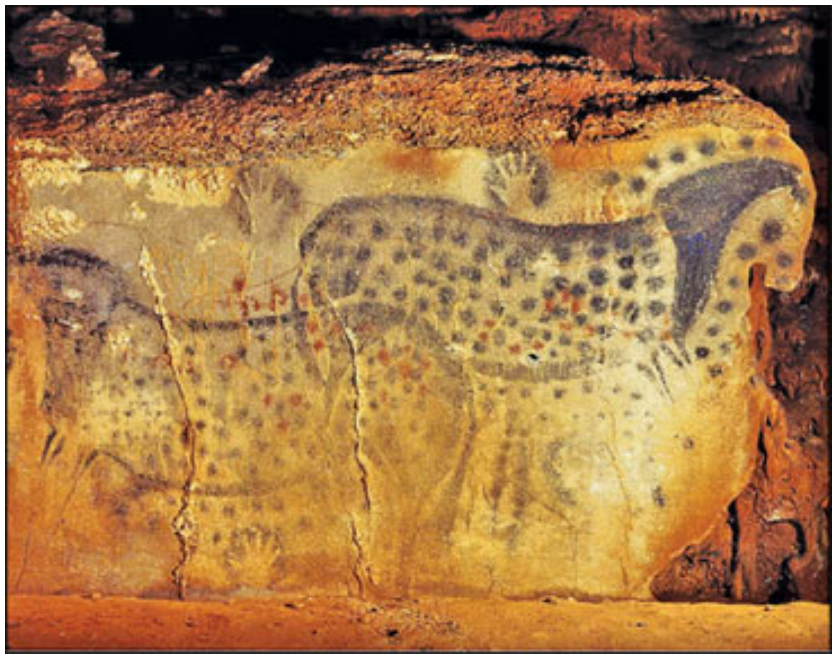


Figure 4.1: The famous spotted horses are found in Pech Merle cave in the Pyrenees of France (From around 25,000 BC). Courtesy of `Sunday Observer`.

geon that during the late 1870s in the Tsukiji Hospital in Tokyo had developed an early method for classification of fingerprints. In 1880 Faulds published his results [59] and wrote a letter to Charles Darwin seeking to collaborate. Darwin, being ill, had no desire for this and referred him to his cousin Francis Galton. Galton showed at this time no interest in the letter from Faulds.

The French anthropologist Alphonse Bertillon devised in 1882 a system for biometric classification called the Bertillon System[1], where photographs and body measurements were recorded, including height, reach, left foot, middle finger etc. As criminals often gave different aliases each time they were arrested, Bertillon tried to solve identification problem via this system. The Bertillon System was quickly adopted and widely used in many countries. In the late 1880s Galton started to doubt whether the Bertillon system was theoretically sound. At this time Galton was pioneering the statistical formulation of correlation and regression. He understood that body measurements are not independent, thereby greatly reducing the number of possible combinations in the Bertillon System. Galton started to look for alternatives to the Bertillon System and began to look into fingerprint recognition. Galton published among other things a book on fingerprints including a new classification system for fingerprints [62]. The unique characteristics of fingerprints identified by Galton are known today as minutiae, but are sometimes referred to as "Galton Details".

In 1903 it was discovered that two persons can be nearly identical in the Bertillon System. A man named Will West was processed into the federal Prison in Leavenworth, Kansas. It was discovered that he had a striking resemblance, see Figure 4.2, and thereby is nearly identical in the Bertillon System, to the inmate known as William West, who was already incarcerated at Leavenworth[2]. After this case, there was a quick decline in the popularity of the Bertillon System, where fingerprint took over as the primary biometric identification system.

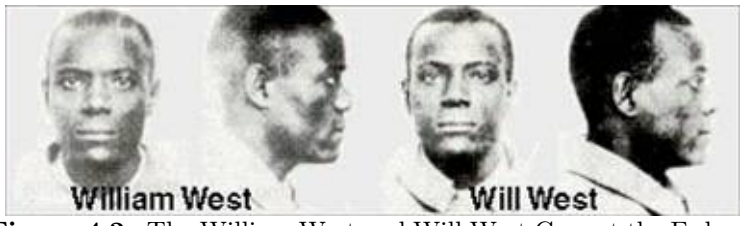


Figure 4.2: The William West and Will West Case at the Federal Prison in Leavenworth, Kansas (1903). Courtesy of `crimescene-forensics.com`.

[1] Also known as "Bertillonage".
[2] It was later discovered that they where identical twin brothers.

Other biometric systems and measures have since emerged, especially after the invention of the computer. The field has grown immensely the last 20 years to the point where we today have facial and fingerprint recognition embedded in commercial computers, phones and cameras. Moreover, it is possible in certain airports to perform check-in via the iris.

Today many countries collect biometric databases and to some extent embed biometric information in passports, where USA is a leading player. However, the largest biometric database in the world today is "Aadhaar", a program under the Unique Identification Authority of India with the ambitious goal of eventually providing reliable national ID documents for most of India's 1.2 billion residents. So far 200 million fingerprint-, face- and iris-biometrics have been recorded with an extra 400 million planned recorded by the end of 2014.

4.2 Elements of Biometrics

What biological measurements qualify as a biometric? Jain *et al.* [89] answers this by stating that any human characteristic can be used as a biometric characteristic as long as it satisfies the following requirements:

Universality each person should have the characteristic.

Distinctiveness any two persons should be sufficiently different in terms of the characteristic.

Permanence the characteristic should be sufficiently invariant (with respect to the matching criterion) over a period of time.

Collectability the characteristic can be measured quantitatively.

In a practical biometric system there are also a number of other issues that should be considered, including:

Performance which refers to the achievable recognition accuracy and speed, the resources required to achieve the desired recognition accuracy and speed, as well as the operational and environmental factors that affect the accuracy and speed.

Acceptability which indicates the extent to which people are willing to accept the use of a particular biometric identifier (characteristic) in their daily lives.

Circumvention which describes the robustness of the biometric characteristic to various fraudulent methods and attacks to the system.

In Table 4.1 a comparison of various biometric characteristics over these requirements can be seen. It is important to stress that the grading is based on the perception of Jain *et al.* [89].

Biometrics	Universality	Distinctiveness	Permanence	Collectability	Performance	Acceptability	Circumvention
DNA	H	H	H	L	H	L	L
Ear	M	M	H	M	M	H	M
Face	H	L	M	H	L	H	H
Facial thermogram	H	H	L	H	M	H	L
Fingerprint	M	H	H	M	H	M	M
Gait	M	L	L	H	L	H	M
Hand geometry	M	M	M	H	M	M	M
Hand vein	M	M	M	M	M	M	L
Iris	H	H	H	M	H	L	L
Keystroke	L	L	L	M	L	M	M
Odor	H	H	H	L	L	M	L
Palmprint	M	H	H	M	H	M	M
Retina	H	H	M	L	H	L	L
Signature	L	L	L	H	L	H	H
Voice	M	L	L	M	L	H	H

Table 4.1: Comparison of various biometric characteristics based on the perception of Jain *et al.* [89]. High, medium and low are denoted by H, M and L, respectively. Redacted from [89].

In biometrics distinction is made between physical, behavioral and adhered human characteristics. These characteristics are in turn grouped into biometrics and soft biometrics, where soft biometrics are characteristics that cannot describe identity single-handedly. A non-exhaustive list of biometric characteristics is given in Table 4.2.

Characteristics type	Biometrics	Soft biometrics
Physical	DNA, ear, face, fingerprint, palmprint, iris, retina, odor, face and hand thermograms, face and hand vein patterns.	gender, height, weight, hand geometry, skin color, eye color, hair color, presence of beard, presence of moustache.
Behavioral	signature, voice.	gait, keystroke.
Adhered human		clothes color, tattoos, branding, accessories.

Table 4.2: Biometric characteristics grouped into physical, behavioral and adhered human characteristics.

4.3 Biometric Recognition Tasks

The three primary biometric recognition tasks are:

- Verification (authentication) - Am I who I say I am? (one to one search)
- Identification (recognition) - Who am I? (one to many search)
- Watch list - Are you looking for me? (one to few search)

Different schemes can be applied to perform the three tasks described above. Which scheme is to be used depends on the nature of the application.

4.3.1 Verification

The verification task is aimed at applications requiring user interaction in the form of an identity claim, e.g., access applications.

The verification test is conducted by dividing persons into two groups:

- **Clients** - people trying to gain access using their own identity.
- **Imposters** - people trying to gain access using a false identity, i.e., an identity known to the system but not belonging to them.

The percentage of imposters gaining access is reported as the False Acceptance Rate (FAR) and the percentage of clients rejected access is reported as the False Rejection Rate (FRR) for a given threshold. An illustration of this is displayed in Figure 4.3.

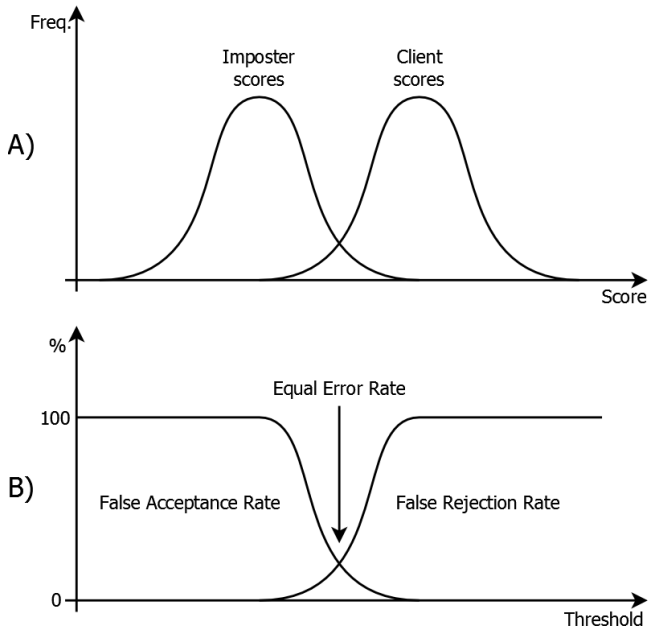


Figure 4.3: Relation of False Acceptance Rate (FAR) and False Rejection Rate (FRR) with the distribution of clients and imposters in a verification scheme. A) Shows the imposters and clients populations in terms of the score (high score means high likelihood of belonging to the client population). B) The associated FAR and FRR curves. The Equal Error Rate (EER) is where the FAR and FRR curves intersect and gives the threshold value for the best separability of the imposter and the client classes.

4.3.2 Identification

The identification task is mostly aimed at applications not requiring user interaction, e.g., surveillance applications.

The identification task works from the assumption that all faces in the test are of known persons. The percentage of correct identifications is then reported as the Correct Identification Rate (CIR), whereas the percentage of false identifications is reported as the False Identification Rate (FIR).

4.3.3 Watch List

The watch list task is a generalization of the identification task which includes unknown people. The watch list task is similarly to the identification test reported in CIR or FIR, but can have FAR and FRR associated with it to describe the sensitivity of the watch list.

4.4 The Process of Biometric Recognition

All different biometric characteristics have the same "process of recognition". This is essentially a pattern recognition problem preformed on the data modality that best captures the biometric characteristic. The "process of recognition" usually[3] consist of four steps, as shown in Figure 4.4: detection (localization), preprocessing, feature extraction and feature matching. These steps are described in the following.

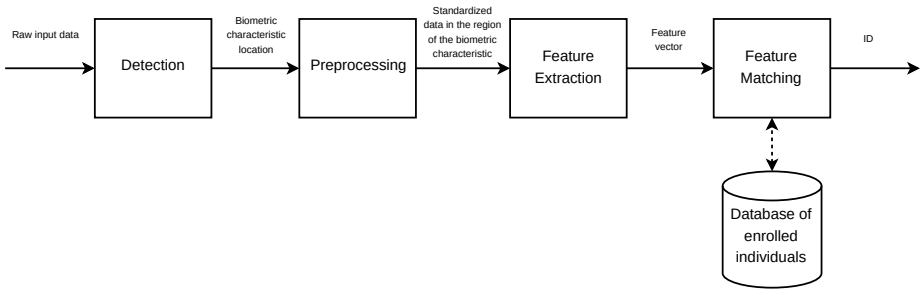


Figure 4.4: The four general steps in biometric recognition.

Detection The aim of the detection step is localization of the biometric characteristic(s) in the data.

[3]Depending on the biometric characteristic changes can be introduced to tailor the performance to a specific application.

Preprocessing The aim of the preprocessing step is to normalize the coarse detection of the biometric characteristic(s), so that a robust feature extraction can be achieved.

Feature Extraction The aim of feature extraction is to extract a compact set of interpersonal discriminating features.

Feature Matching Feature matching is the actual recognition process. The feature vector obtained from the feature extraction is matched to classes (such as individuals, gender, etc.) already enrolled in a database. The matching algorithms vary from the fairly simple Nearest Neighbor to advanced schemes such as Neural Networks.

4.5 Biometric Recognition Considerations

In any biometric system a number of common considerations have to be addressed. In the following some of these are described. Within this work the "Curse of dimensionality", described in Section 4.5.2, is considered as one of the most important among them.

4.5.1 Strength of the Biometric Feature

An initial overview of the strength of different biometric characteristics has already been given in Table 4.1. However, depending on the application at hand some issues need to be addressed with respect to the strength of the system.

4.5.1.1 Biometric Spoofing

Biometric spoofing is the act of a malicious imposter (actively trying to circumvent the system), trying to defeat the biometric system through the introduction of fake biometric samples. This is usually performed by artificially created biometric characteristic such as: image or video recordings of a face or iris, artificial rubber fingers, voice recordings, etc. Ways of limiting such spoofing include:

- Liveness detection based on recognition of physiological activities as signs of life (head/eye/pupil movement, eye blinking, finger perspiration, etc.).
- Introducing challenge-response mechanisms that are difficult to pre-record.
- Human supervision of the biometric verification process.

4.5.1.2 Multi-modal Biometrics

Some biometrics are easily copied (face, iris, voice, fingerprint, etc.). Here multi modal biometrics are very good at limiting the possibility for impersonating other persons, by requiring more than one biometric characteristic to accept an ID-verification. Even soft biometrics can be included into the system, such as in [130, 133].

4.5.2 High Dimensional Input Space and Small Sample Size

A problem associated with biometric recognition is the high dimensional input space (2D/3D image domain, DNA sequence, etc.) and the usually small sample size of individuals. A DNA sequence has in total 3 billion highly correlated data variables, a 2D image consisting of HD quality (1920 $\times$ 1080 pixels) can be treated as a point that resides in a 2,077,600-dimensional space.

This problem has led to the introduction of the term "Curse of dimensionality", which was first coined by Richard Bellman in 1961 [13] by this quote:

> *In view of all that we have said in the forgoing sections, the many obstacles we appear to have surmounted, what casts the pall over our victory celebration? It is the curse of dimensionality, a malediction that has plagued the scientist from the earliest days.*

In [145] Steinbach *et al.* note that:

> *The issue referred to in Bellman's quote is the impossibility of optimizing a function of many variables by a brute force search on a discrete multidimensional grid. (The number of grids points increases exponentially with dimensionality, i.e., with the number of variables.) With the passage of time, the "curse of dimensionality"*

has come to refer to any problem in data analysis that results from a large number of variables (attributes).

This means that today the "Curse of dimensionality" also refers to the issue caused by the rapid increase in "volume" associated with adding extra dimensions to a space. It follows that in a high dimensional space data will be sparse, as it is not feasible to add data samples at an exponential rate.

In classification problems, where we want to "learn the nature of" (predict) our data from a finite number of data samples in a high dimensional space, an enormous amount of training data is required to ensure that we sample the distribution of data correctly. This gives rise to the Hughes phenomenon or Hughes effect [85], which explains that with a fixed number of training samples the predictive power reduces as the dimensionality increases, as seen in Figure 4.5.

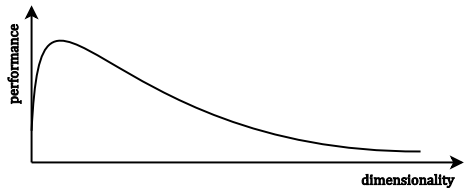


Figure 4.5: The Hughes phenomenon: with a fixed number of training samples the predictive power reduces as the dimensionality increases.

A different way to look at this is the following. When the dimensionally is growing the data samples density will become more and more sparse. This is easily understood by the following example. Consider that we want to uniformly sample the unit line 10 times. If we add another dimension, then we need to sample 10^2 times to perform the same dense sampling of the unit square. In the unit cube (3D) this grows to 10^3 and so forth. If we do not add data samples, then the data sample density must go down (become more sparse) at an exponential rate.

Another very interesting property of high dimensional space is that the Euclidian distance and data placement become "counter intuitive". Consider the

equation for the volume of a hyper-sphere [70]

$$V_n(r) = \frac{2r^n \pi^{n/2}}{n\Gamma(\frac{n}{2})}, \tag{4.1}$$

where Γ is the Gamma function

$$\Gamma(z) = \int_0^\infty t^{z-1} e^{-t} dt. \tag{4.2}$$

Imagine we place two hyper-spheres A and B with the same center with radius a and b, where $b = a + \xi$ (a 2D representation is shown in Figure 4.6).

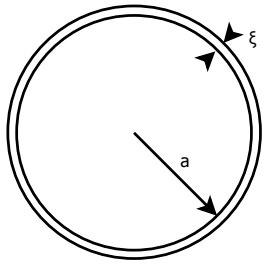


Figure 4.6: Two hyper-spheres with the same center with radius a and $b = a + \xi$.

The volume of the crust "ξ" is now defined as

$$V_n(crust) = \left(1 - \frac{V_n(a)}{V_n(b)}\right) V_n(b). \tag{4.3}$$

The volume ratio of the two hyper-spheres A and B can be simplified in this way

$$\frac{V_n(a)}{V_n(b)} = \frac{\frac{2(b-\xi)^n \pi^{n/2}}{n\Gamma(\frac{n}{2})}}{\frac{2b^n \pi^{n/2}}{n\Gamma(\frac{n}{2})}} = \frac{(b-\xi)^n}{b^n} = \frac{b^n \left(1 - \frac{\xi}{b}\right)^n}{b^n} = \left(1 - \frac{\xi}{b}\right)^n. \tag{4.4}$$

It is now clear that even if ξ is very small, the ratio will go towards zero in a high dimensional space. Then from Equation 4.3 follows the conclusion: "All the volume is in the crust!"[4]. This indicates that a local neighborhood in higher dimensions is no longer local. Beyer *et al.* [14] also concludes that under certain conditions the distances between a data point and the rest of the date points will be the same in high dimensional space leaving distance algorithms such as the Nearest Neighbor very ineffective.

One of the most used methods to counter the effect of the "Curse of dimensionality" is dimensionality reduction. As noted by [26, 28, 165], for most practical applications, data variables in a high dimensional input space will be (highly) correlated. In this case each input variable contributes only little (or nothing) to the structure of the data that is interesting for the given application. Thus, the problem of dimensionality reduction is to find a mapping from the high dimensional input space to a lower, compact subspace given by

$$F(x) : \mathbb{R}^h \rightarrow \mathbb{R}^n \tag{4.5}$$

where $n \ll h$. In this subspace the problems described above will be much less prominent.

It should also be noted that in some respects there is a counterpart conception of "Blessing of dimensionality", that refers to some of the positive elements of going towards high dimensional space. This is beyond the scope of this thesis. A good starting point for further reading is the overview given by Donoho in [43].

4.6 Discussion

How to exactly tailor the biometric recognition pipeline is determinated by the specifications for the application or device it is being developed for. Questions such as "do we expect malicious users?" or "is the application to be used in a controlled environment?" should be thoroughly examined to design the application which will result in the best possible performance. A recent example of a consumer biometric recognition application is the new iPhone TouchID fingerprint recognition. The security of this system was broken (spoofed) within 48 hours of the new iPhone release, indicating that this system should not

[4]It can be noted that "All the volume of the crust" also can be infinitely small when going towards very high dimensions.

guard access to safes or doors. Nevertheless, it is still safer than its four digit predecessor, assumably faster and more user-friendly. This shows that the application's/device's security, user-friendliness, etc. should be tailored to the user requirements.

In this work we will look only at the core of the biometric recognition process, not taking into account specification issues such as malicious users, user-friendliness, etc. Special emphasis will be put on dimension reduction / feature extraction in the subsequent chapters.

Chapter 5

Methods

In this chapter core methods used in the contributions in this biometric part of the thesis are presented. The methods and contributions from the biometric part of this thesis deal with the last two steps in the biometric recognition process, described in Section 4.4, namely Dimension Reduction (Feature Extraction) and Classification (Feature Matching).

5.1 Dimension Reduction

In most practical cases a high dimensional data set will reside on a lower dimensionality manifold with respect to the problem/application at hand. This embedded manifold has dimensionality d and is called the data set's intrinsic dimensionality. Unfortunately, neither the intrinsic dimensionality nor the geometry of this embedded manifold is known, which makes dimensionality reduction an ill-posed problem that can only be solved by assuming certain properties of the data. Dimension reduction is therefore the task of learning the underling manifold. When the manifold is found it can be used for different purposes such as:

- Low dimension data representation: In the case of having a high dimen-

sionality data sample a compression of the data can be performed with little or no loss.

- Visualizing the intrinsic degree of freedom of the data: As it is shown in the contributions in Appendix B and C, it is possible to visualize what is captured in one or more dimensions on the manifold in the original input space.
- Data preprocessing: As discussed in Section 4.5.2 many methods perform better in lower dimensional space and dimensionality reduction is widely used in machine learning and pattern recognition algorithms.

Many different dimensionality reduction techniques has been proposed in the literature. These techniques are normally grouped into linear or non-linear. Sometimes they are also grouped into whether they preserve the global or the local structure of the data. In Table 5.1 selected dimensionality reduction techniques are shown.

Preserving	Technique	Method
Global Structure	Linear	Linear Discriminant Analysis
		Principal Component Analysis
	Non-linear	Kernel Linear Discriminant Analysis
		Kernel Principal Component Analysis
Local Structure	Linear	Locality Preserving Projections
	Non-linear	Isomap
		Laplacian Eigenmap

Table 5.1: Dimensionality reduction methods.

In the following the different dimensionality reduction techniques used in this work are presented. It can be noted that Linear Discriminant Analysis can be thought of both as a dimensionality reduction method and as a classification method. For the techniques not used in this work we direct the reader to [23, 65, 153] for a good overview.

5.1.1 Principal Component Analysis

Principal Component Analysis (PCA), also known as Karhunen-Loève transformation, is a linear transformation which captures the variance of the input data.

The coordinate system in which the data resides is rotated by PCA, so that the first-axis is parallel to the highest variance in the data (in a one-dimension projection). The remaining axes can be explained one at the time as being parallel to the highest variance of the data, while all axes are constrained to be orthogonal to all previous found axes. PCA is an unsupervised method that is widely used in data analysis, where the first-axis will contain the highest variance, the second-axis contains the second highest variance and so on. An example in two dimensions is shown in Figure 5.1.

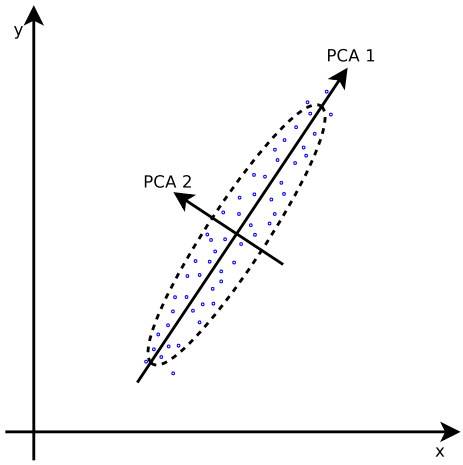


Figure 5.1: An example of PCA in two dimensions, showing the PCA axis that maximizes the variation in the first principal component: PCA 1.

The basis axes of PCA are given by the eigenvectors of the covariance matrix of the data[1]

$$\mathbf{\Sigma}_{\hat{\mathbf{X}}}\mathbf{\Phi}_{\hat{\mathbf{X}}} = \mathbf{\Phi}_{\hat{\mathbf{X}}}\mathbf{\Lambda}_{\hat{\mathbf{X}}}, \tag{5.1}$$

where $\mathbf{\Sigma}_{\hat{\mathbf{X}}}$ is the covariance matrix, $\mathbf{\Phi}_{\hat{\mathbf{X}}}$ the eigenvectors and $\mathbf{\Lambda}_{\hat{\mathbf{X}}}$ the diagonal matrix containing the corresponding eigenvalues. A more comprehensive overview of PCA and how to calculate it has been provided by the author of this thesis in [50], where the computational issue of PCA in a high dimensional space of dimension k is addressed. The computational issue of PCA arises from the fact that PCA is calculated from the covariance matrix that is a square

[1] More efficient algorithms exist to calculate the PCA from the singular value decomposition. However, these will not be discussed further in this work.

matrix with size $k \times k$. This problem is very hard to solve when k gets big. However, it is possible to calculate the eigenvectors of the non-zero eigenvalues from a much smaller[2] matrix with size $n \times n$, where n is the number of samples in the data set [36].

5.1.2 Sparse Principal Component Analysis

Sparse Principal Component Analysis (SPCA) is an extension to PCA where the new basis is sought sparse (contains many zeros in the SPCA basis vectors) by sacrificing basis orthogonality. Depending on how the sparse basis vectors are obtained these can be more or less orthogonal. There are different ways to formulate the SPCA objective function [90, 171], where the simplest is thresholding the original PCA basis vectors and setting small contributions (under the threshold) to zero in the PCA basis vectors. In this work we use the generalized power method described in [90] with objective function

$$\phi_{\ell_0}(\gamma) = \max_{\mathbf{x} \in S^P} \sum_{i=1}^{n} [(\mathbf{a}_i^T \mathbf{x})^2 - \gamma]_+, \tag{5.2}$$

where $[\ldots]_+$ truncates all negative values to zero, S^P refers to a point on the unit hypersphere, n is the number of variables in each sample, and γ is the sparsity parameter. The γ parameter controls the sparsity of the SPCA solution between the two extremes: a full PCA and the zero solution. It depends on the application whether sparsity or explained variance is preferred.

Equation 5.2 yields a single Sparse Principal Component (SPC). In order to obtain more than one SPC, (5.2) is used iteratively by means of deflation. Deflation is obtained as follows. When a new SPC is obtained its variation is removed from the sample variance before obtaining the next SPC. The advantage of this is that the resulting SPC number k will be nearly orthogonal to the $k-1$ previously computed SPCs.

5.1.3 Other Methods

During this work the following other methods for representing facial data have been investigated: Independent Component Analysis (ICA), Archetype Analysis (AA), Non-Negative Matrix Factorization (NMF) and Minimum/ Maximum

[2] Assuming we have less number of samples in the data set than the dimension of the data.

Autocorrelation Factor analysis (MAF). All of these methods are well established and have multiple uses. However, none of these methods was used in the contributions produced as part of this thesis, as it was found that they produced similar or worse results compared to PCA. PCA appears to be very robust when seeking to represent data in a compact format for further investigation by other algorithms. The fact that PCA in this work preformed best should not be taken as it is better than the other algorithms, since its performance is very application-specific and in other setups PCA might not be the best choice. A short description of the four algorithms is listed in the following:

ICA finds a statistically independent basis. This makes it ideal for unmixing a mixed source, such as the well known "cocktail party problem".

AA makes a new basis by representing each individual data sample as a mixture of archetypes. The resulting archetypes are the extreme data samples that encapsulate the data set in a convex hull.

NMF produces a basis which compared to most other methods does respect the non-negativity structure of the original data. Furthermore, NMF produces a basis that represents local components of the original data.

MAF takes into account the spatial nature of data. The MAF transform minimizes the autocorrelation so that the resulting basis is ordered by the minimum autocorrelation between neighboring pixels.

5.2 Classification

Classification is the task of assigning the correct label/class to an unknown sample. In this work two classification algorithms are used namely Linear Discriminant Analysis (LDA) and Support Vector Machine (SVM).

5.2.1 Linear Discriminant Analysis

Linear Discriminant Analysis is a linear transformation. It is a supervised method, which implies that all training-data samples must be associated (manually) with a class. LDA maximizes the between-class variance as well as minimizes the within-class variance. A 2D example of LDA is shown in Figure 5.2.

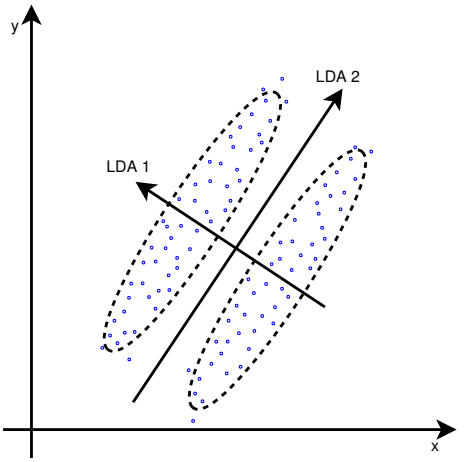


Figure 5.2: An example of LDA in two dimensions, showing the LDA axis that maximizes the separation between the classes and minimizes the variation inside the classes.

LDA uses the Fisher criterion that minimizes the within-class variance while maximizing the between-class variance. The objective function for LDA is

$$\max_{\mathbf{W}} \frac{\mathbf{W}^T \mathbf{S}_B \mathbf{W}}{\mathbf{W}^T \mathbf{S}_W \mathbf{W}}, \tag{5.3}$$

where the $\mathbf{S}_B$ is a matrix describing the between-class variance and the $\mathbf{S}_W$ is a matrix describing the within-class variance.

In biometric applications $\mathbf{S}_W$ is nearly always singular. This is due to the fact that the rank of $\mathbf{S}_W$ is at most $n - c$ (the number of training samples minus the number of classes), where n usually is much smaller than the number of pixels in each image. In order to overcome that $\mathbf{S}_W$ is singular a PCA is usually performed[3] on the images prior to LDA, which removes redundancy and makes the data samples more compact.

A more comprehensive overview of LDA has been provided by the author of this work in [50].

[3]Normally capturing between 95% to 99% of the variance.

5.2.2 Support Vector Machine

Support Vector Machine is a binary classifier. Similar to LDA it is a supervised method that requires labeled training data to find a decision boundary. SVM maximizes the margin of the decision boundary. A 2D example of a SVM is shown in Figure 5.3, where a hyperplane that separates the classes exists.

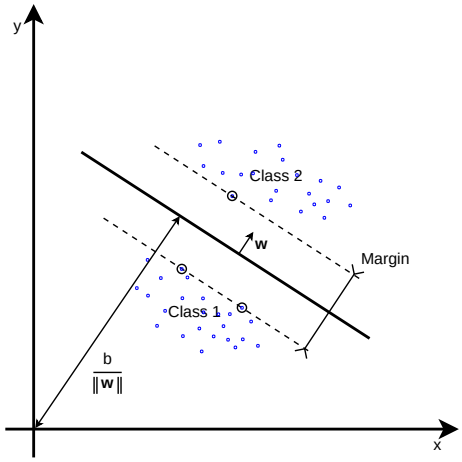


Figure 5.3: An example of SVM in two dimensions showing the SVM hyperplane that maximizes the margin between the classes. The circled data points are the support vectors.

The hyper plane that separates the classes can be formulated as

$$\mathbf{x}\mathbf{w} + b = 0, \tag{5.4}$$

where $\mathbf{w}$ is normal to the hyperplane and $\frac{b}{\|\mathbf{w}\|}$ is the perpendicular distance from the hyperplane to the origin. The two hyperplanes that make up the margin between the two classes can be formulated as

$$\mathbf{x}_i\mathbf{w} + b \leq 1 \texttt{ for } \mathbf{x}_i \in \textit{class one}, \tag{5.5}$$

$$\mathbf{x}_i\mathbf{w} + b \geq 1 \texttt{ for } \mathbf{x}_i \in \textit{class two}. \tag{5.6}$$

It is clear that the two hyperplanes lie $\frac{|1-b|}{\|\mathbf{w}\|}$ and $\frac{|-1-b|}{\|\mathbf{w}\|}$ from the origin. This results in a margin with a width of $\frac{2}{\|\mathbf{w}\|}$. To maximize the width of the margin $\|\mathbf{w}\|$ must be minimized, giving the objective function of SVM

$$\min \|\mathbf{w}\| \; s.t. \; y_i(\mathbf{x}_i\mathbf{w} + b) - 1 \geq 0 \; \forall_i, \tag{5.7}$$

where the last term is a combination of equation 5.5 and 5.6, y_i is the class label so that $y_i \in \{-1, 1\}$. This can be turned into a convex quadratic optimization problem with the use of Lagrange multipliers. SVM can also be extended with the concept of "soft margins" that allows for the case where there are no hyperplanes that can perfectly separate the data. For a more in-depth description of SVM the author recommends [22, 78] for further reading.

5.3 Discussion

Many of the methods described in this chapter can be extended to a non-linear form by the "kernel trick" [3]. The kernel trick is a way of mapping data from the input space into a very high dimensional space (possibly infinite) and down to an inner product space without having to compute the mapping to the very high dimensional space explicitly. This gives us the possibility to use linear methods that will behave similar to non-linear methods in the original input space. An overview of how to use the kernel trick on LDA has been provided by the author of this thesis in [50]. It is important to note that the use of the "kernel trick" requires parameter tuning, and finding the optimal parameters is not a trivial problem to solve. Non-linear methods can also lead to overfitting when the number of samples is small compared to the input space. In the contributions in this work both linear and non-linear methods are used with an emphasis on the linear methods, as these require less or no tuning and are easier to compare. Furthermore, for most cases it is impossible to prove that the tuning parameters of a non-linear method have been chosen optimally.

Depending on the data, the differences between the solutions of LDA and SVM can be very subtle. This effect can be seen in contribution [B] where the performance is very similar. However, this is not always the case and even though SVM is among the most popular classifier algorithm at the moment in the literature, it will not always be the optimal solution to a problem. Which classifier or kernel should be employed depends greatly on the nature of the data.

Part II

Facial Analysis in a Genome-Wide Context

Chapter 6

Human Genetics

This chapter presents a general introduction into specific concepts within the field of *Human Genetics* that are used further in this work. The concepts are presented in an easy-digestible form with focus on their relation to facial morphology. For a more thorough description of the field of *Human Genetics* we direct the reader to ENCODE, the Encyclopedia of DNA Elements, that is a project funded by the National Human Genome Research Institute to identify all regions of transcription, transcription factor association, chromatin structure and histone modification in the human genome sequence. The information is easily accessible trough the Nature ENCODE explorer at
`http://www.nature.com/encode/`.

6.1 Introduction to Human Genetics

The human genome is the complete set of human genetic information which is copied and inherited across generations. In the nucleus of every human cell is a copy of the genome. The entirety of our hereditary information is stored in the genome. The genome consists of 23 pairs of chromosomes (44 autosomes and 2 sex-chromosomes). The chromosomes are numbered from 1 to 22 according to size, where chromosome 1 is the longest. The sex-chromosomes are labeled X and Y, where women have two identical X chromosomes and men have one

X and one Y chromosome. Every chromosome consists of one long polymer of deoxyribonucleic acid (DNA) in the well known double helix structure, as seen on Figure 6.1. The chromosomes consist of two arms (parts), a short and a long arm, denoted p and q respectively, where the DNA polymer is tightly packed. The genetic information of DNA is encoded along the double helix structure via four nucleotides: guanine, adenine, thymine, and cytosine. The elements of this encoding are called a base pair and have to be one of four configurations: guanine-cytosine, cytosine-guanine, adenine-thymine and thymine-adenine as seen on Figure 6.2.

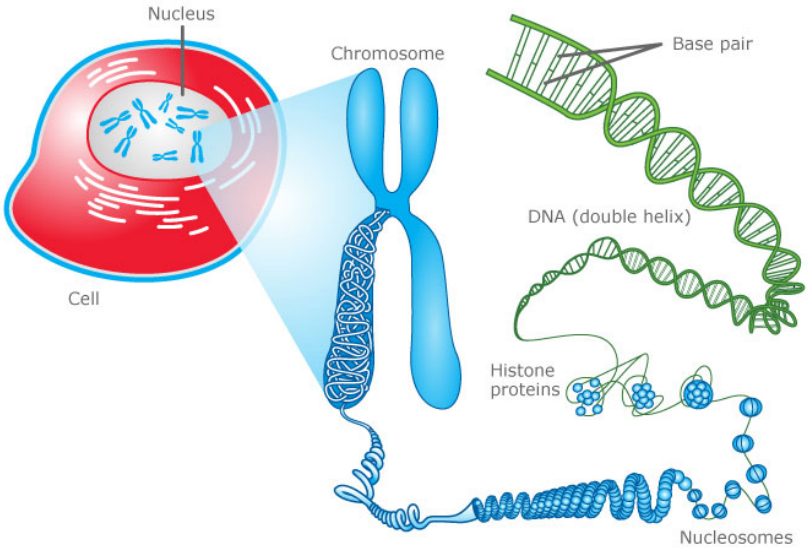


Figure 6.1: Overview of the human cell nucleus, chromosome structure and DNA. Courtesy of `The University of Waikato`.

The information stored along the DNA string is organized into gene and non-gene sequences, where the gene sequences are further organized into exon and intron sequences. The exon sequences are the part of the gene that directly codes for proteins, where the intron sequences do not. The coding of proteins is how the genome "controls" the behavior of a particular cell. We understand much more about how an exon affects the behavior of a cell, than how an intron does, and even less about how the non-gene sequences affect the behavior. This does not mean that the intron and non-gene sequences are not important, but our understanding of them is considerably limited.

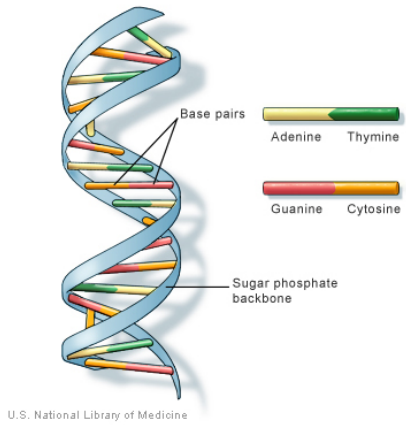


Figure 6.2: The base pair structure of DNA. Courtesy of `U.S. National Library of Medicine`.

DNA was first identified and isolated by Friedrich Miescher in 1869 and the double helix structure of DNA was first discovered by James Watson and Francis Crick in 1953 [161] with the use of the "photo 51" closeup x-ray image of DNA produced by Raymond Gosling and Rosalind Franklin in 1952 (see Figure 6.3).

6.2 Single-Nucleotide Polymorphism

In the human genome there are about $32 \cdot 10^3$ genes, where more undiscovered genes probably exist, and about $3 \cdot 10^9$ base pairs. It is estimated that above 99%[1] [129] of the base pairs are exactly the same in all humans, regardless of race or heritage. Disregarding the copy-number variation, explained in the next section, the remaining of the base-pairs in the human genome is denoted single-nucleotide polymorphism (SNP). A SNP is a DNA sequence variation occurring in a single base pair over the human population, as shown in Figure 6.4.

Most SNP locations have two variations (as seen in Figure 6.4), but can have up to four. These variations are named alleles. Over the entire human population (or geographical or ethnic group) the minor allele frequency (MAF) can be

[1]Depending on the particular study, numbers between 99.6% and 99.9% are reported. However, these studies disregard very rare variation below a threshold in the range 1%-5%.

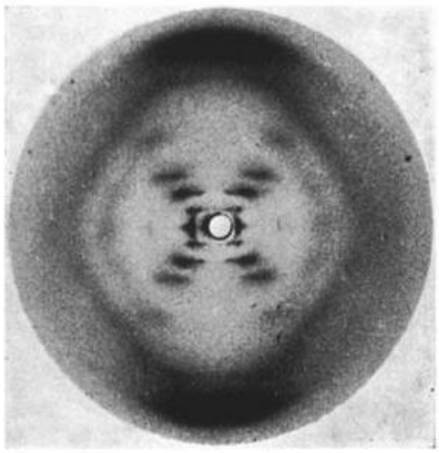


Figure 6.3: The "photo 51" closeup x-ray image of DNA produced by Raymond Gosling and Rosalind Franklin. Courtesy of `bbc.co.uk`.

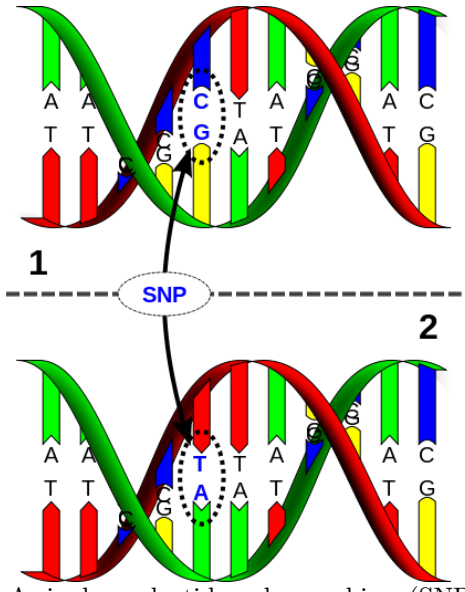


Figure 6.4: A single-nucleotide polymorphism (SNP), a DNA sequence variation occurring in a single base pair over the human population. Courtesy of `David Hall`.

estimated for a given SNP. MAF is the lowest allele frequency at the position of the SNP that is observed in the population in question. The MAF will vary for some SNPs between geographical or ethnic groups.

6.3 Copy-Number Variation

Copy-number variation (CNV) is a large structural variation of the DNA. CNV includes deletions, duplications and inversions of the DNA as shown in Figure 6.5. The size of these mutations range from about one kilobase (kb), 1000 base pairs, to several megabases (mb) [126]. Even though CNVs are much more rare than SNPs, it has been discovered that CNVs are not uncommon occurrence in the human genome [87, 137]. A specific CNV (specific size and place in the DNA) can however be very rare. CNVs have been associated with a higher or lower risk of getting a certain disease. For example, in [31, 143] an association between CNVs and schizophrenia has been shown, which will be investigated further in this thesis.

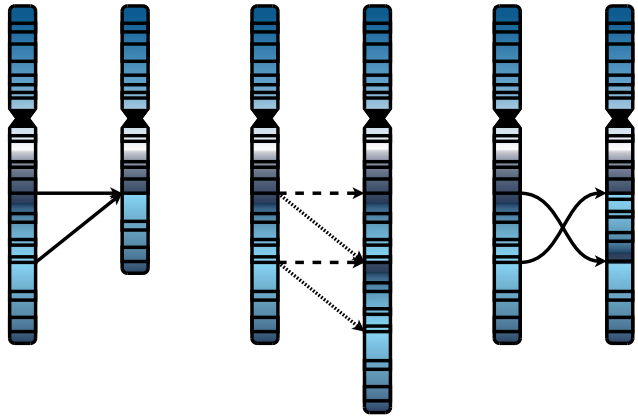


Figure 6.5: Copy-number variation, showing from the left deletions, duplications and inversions.

6.4 Heritability

The facial appearance of an individual is due to the combinatorial effect of genetics, the environment, and possibly some other factors. Currently, it is unknown

how much of the facial appearance is due to the human genome. Furthermore, which parts of the genome contribute to specific facial development is also largely unknown. Only recently have genome-wide association studies (GWASs) uncovered a few SNPs contributing a little to specific facial distances [102, 119]. We will discuss GWAS in detail in Section 7.1. Complex facial development due to the genome is unknown and no findings have been presented in this area.

Beyond facial features, schizophrenia is also highly hereditary. Furthermore, schizophrenia appears to be associated to minor physical anomalies (MPA) in the face, which will be elaborated on further in this work. In the following a short overview of schizophrenia is given.

6.4.1 Schizophrenia

The word schizophrenia is derived from the two Greek words "skhizein" that translates into 'to split' and "phren" that translates into 'mind'. However, contrary to some public perception, schizophrenia does not imply a "split mind or personality" ("multiple personality disorder" or "dissociative identity disorder"). Schizophrenia is a "splitting of mental functions" [9].

Schizophrenia is a mysterious mental disorder, where current treatments provide a form of control rather than a cure with lifelong treatment with antipsychotic medication. It mainly affects cognition, but also chronic problems with behavior and emotion can be contributed to schizophrenia [155]. Men and women from all cultural backgrounds are affected equally, and the lifetime risk of getting schizophrenia is approximately one percent for the population worldwide.

Schizophrenia is a spectrum of disorders with a range of different symptoms. The ICD-10 [1] published by the World Health Organization classifies schizophrenia into nine disorders: Paranoid schizophrenia, Hebephrenic schizophrenia, Catatonic schizophrenia, Undifferentiated schizophrenia, Post-schizophrenic depression, Residual schizophrenia, Simple schizophrenia, Other schizophrenia, and Schizophrenia unspecified.

Schizophrenia has no precise known cause and diagnosing schizophrenia is based on the presence of typical signs and symptoms. The level of severity and the number of symptoms vary among the patients [141].

6.4.2 CNV and Facial Features Correlation to Severe Schizophrenic Disorders

Lichtenstein *et al.* [101] reported that the chance of getting schizophrenia is increased if schizophrenia was present in the family. Figure 6.6 presents a selective overview of the heredity of schizophrenia. In Stefansson *et al.* [143] association between specific CNV and schizophrenia is found.

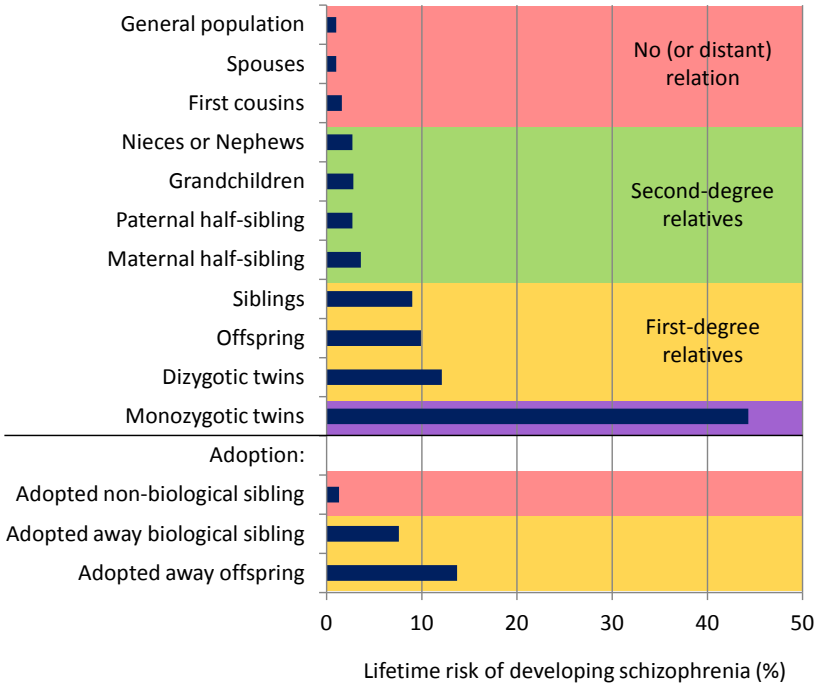


Figure 6.6: Heredity of schizophrenia, redacted from [101, 109].

Prior studies have also suggested a connection between MPA and psychiatric disorders especially within the schizophrenic spectrum, where the MPA were located in the face and on the hands in particular [162].

Chapter 7

Methods

In this chapter core methods used in the contributions in this genetic part of the thesis are presented.

7.1 Genome-Wide Association Studies

Genome-wide association study (GWAS) is the study of many (all) genetic variants in different individuals to see if any variant is associated with a trait. GWAS typically focuses on associations between SNPs and characteristics or traits or diseases, also know as a phenotype. GWAS is usually performed in a patient-control or a quantitative phenotype setup. A patient-control setup is where a sample of individuals, also called "cohort", is tested for a difference in allele frequency between the patient and the control cohort. In the quantitative GWAS setup a quantitative phenotype such as "human height" is tested for correlation via a regression. In both types of GWAS co-variates can be included to account for variance in the phenotype for traits such as sex, age, etc. To conduct a GWAS the cohort is first genotyped for known SNPs. Then, the single SNPs are associated to the phenotype and co-variates. The number of genotyped SNPs is constantly rising as better and cheaper technology becomes available. At time of writing, genotype chips of 500,000 to 1,000,000 SNPs are common. The genotype chip, also called a DNA microarray chip, is the most

used form to retrieve large scale SNP information from an individual. In short, the process of retrieving an individual's SNP information is done by chemically breaking the individual's DNA into tiny pieces, residing in a liquid. This liquid is poured on top of the genotype chip. These chips are covered with hundreds of thousands of little testing probes. The probes are made up of molecules to which the exact matching pieces of DNA naturally attach. These molecules are designed to chemically light up when a match occurs. In this way the SNP information is retrieved in parallel on the chip. An example of an Illumina genotype chip is shown in Figure 7.1. Interactions between SNPs are not taken into account in a GWAS. GWAS is usually reported in effect size and a p-value for the significance of the test.

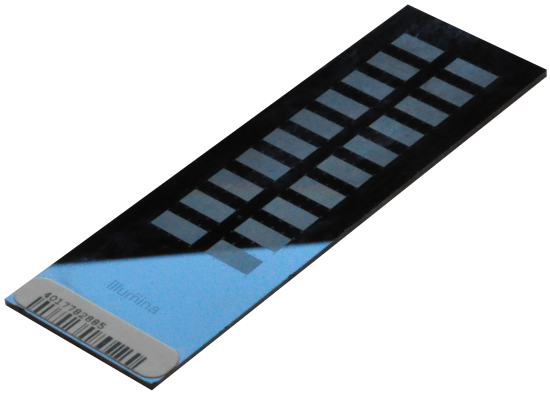


Figure 7.1: Illumina genotype chip.

When doing GWAS the issue of multiple testing must be addressed. A traditional method for accounting for multiple tests is the Bonforonni correction, where the original significance threshold, normally 0.05, is lowered by dividing by the number of performed tests. A Bonforonni correction is typically considered conservative if the tests are related. In GWAS a Bonforonni correction is considered to be too conservative of a threshold, as SNPs located closely together on the DNA will be highly correlated. Thus, a more relaxed correction conditions are used with GWAS, but will vary depending on the GWAS setup and geographical and ethnical background of the individual participants.

The GWAS in this work was conducted at deCODE Genetics located in Iceland. The method and setup used at deCODE are such that the p-value must be below 10^{-8} to be considered genome-wide significant. Furthermore, in GWAS in general and also at deCODE it is favored to perform experiments in a two stage setup where first a discovery is found in a cohort retaining a p-value below 10^{-8} and then this signal is replicated in a replication (unseen) cohort retaining

a p-value below 0.05.

The underlying rationale for GWAS is the "common disease, common variant" hypothesis. It states that common disease alleles underlie most common diseases (probably in connection with environmental factors) [128, 160]. However, common variants appear only in rare cases to have a high explanatory effect. Figure 7.2 shows the feasibility of identifying genetic variants by risk allele frequency and strength of genetic effect. The effect size (odds ratio) is described in more detail in Section 7.2.4. Most research work in the field has found associations which lie between the dotted lines, in particular in the upper left and the lower right corners of the graph. Researchers speculate that there is much more to be discovered in the "center", with low allele frequency and moderate effect size. Even though researchers have worked extensively with GWAS during the last decade, our knowledge only scratches the surface of the secrets of the human genome. For example, at least 40 SNPs have been associated with human height, with an estimated heritability of about 80%. Yet they explain only about 5% of the phenotypic variance despite studies of tens of thousands of people [107].

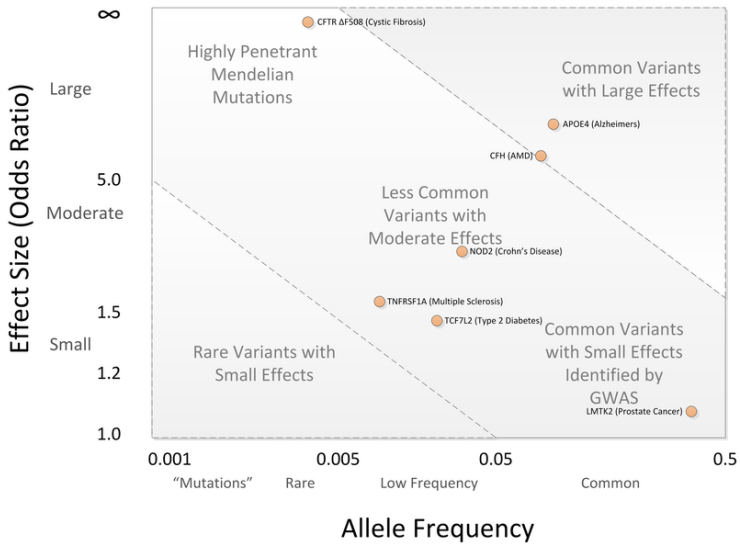


Figure 7.2: Spectrum of Allele Effects. Most research results are concentrated in between the dotted lines, in particular in the upper left and the lower right corner. Courtesy of `Bush et al.` [25].

7.2 Two-sample Statistic

Most of the studies in this work are done on two-sample phenotypes, patients and controls. In this section a short description of the methods for calculating if there is a statistically significant difference between the two samples (patient/control) is presented. The different methods are referred to as "parametric" or "non-parametric". The parametric statistical test makes assumptions about the samples distribution, while the non-parametric test is one that makes no such assumptions, thus having the drawback of lower statistical power.

7.2.1 Two-sample T-test

The two-sample T-test, also called Student's T-test, is a parametric test. It tests the null hypothesis against the alternative hypothesis. The null hypothesis states that the data are independent random samples from a normal distribution with equal means and equal but unknown variances, whereas the alternative hypothesis states that the means are not equal. The T-test statistic is

$$T = \frac{\overline{\mathbf{x}} - \overline{\mathbf{y}}}{s_{xy}\sqrt{\frac{1}{n_x} + \frac{1}{n_y}}}, \tag{7.1}$$

where $\overline{\mathbf{x}}$ and $\overline{\mathbf{y}}$ are the samples means, n_x and n_y are the sizes of the two samples and s_{xy} is the pooled sample standard deviation expressed by

$$s_{xy} = \sqrt{\frac{(n_x - 1)s_x^2 + (n_y - 1)s_y^2}{n_x + n_y - 2}}, \tag{7.2}$$

where s_x^2 and s_y^2 are the two samples' squared standard deviations. The T-test statistic can then be related to the t-distribution with $n_x + n_y - 2$ degrees of freedom and a p-value for statistical significance can be obtained.

7.2.2 Wilcoxon Rank Sum Test

Wilcoxon rank sum test, also known as the Mann–Whitney U test, is the non-parametric counterpart to the two-sample T-test. Instead of comparing two

population means, the two population medians are compared. The U statistic is given by

$$U = \min\left(r_x - \frac{n_x(n_x+1)}{2}, r_y - \frac{n_y(n_y+1)}{2}\right), \tag{7.3}$$

where r_x and r_y are the sum of the ranks of the "scalar values" for the two samples, respectively.

For large samples[1], the U statistic can be related to the normal distribution and a p-value for statistical significance can be obtained by

$$\frac{U - \mu_U}{\sigma_U}, \tag{7.4}$$

where

$$\mu_U = \frac{n_x n_y}{2} \tag{7.5}$$

and

$$\sigma_U = \sqrt{\frac{n_x n_y (n_x + n_y + 1)}{12}}. \tag{7.6}$$

7.2.3 Hotelling's Two-sample T-squared Test

Hotelling's two-sample T-squared test is a multivariate equivalent of the two-sample T-test. The Hotelling's two-sample T-squared statistic is given by

$$T^2 = \frac{n_x n_y}{n_x + n_y}(\overline{\mathbf{x}} - \overline{\mathbf{y}})^T S^{-1}(\overline{\mathbf{x}} - \overline{\mathbf{y}}), \tag{7.7}$$

[1]When n_x and n_y both are greater than eight. For smaller samples special lookup tables must be used.

where $\overline{\mathbf{x}}$ and $\overline{\mathbf{y}}$ are the samples mean, n_x and n_y are the sizes of both samples respectively, and the unbiased pooled covariance matrix estimate is

$$S = \frac{\sum_{i=1}^{n_x}(\mathbf{x}_i - \overline{\mathbf{x}})(\mathbf{x}_i - \overline{\mathbf{x}})^T + \sum_{i=1}^{n_y}(\mathbf{y}_i - \overline{\mathbf{y}})(\mathbf{y}_i - \overline{\mathbf{y}})^T}{n_x + n_y - 2}. \tag{7.8}$$

The Hotelling's two-sample T-squared statistic can then be related to the f-distribution and a p-value for statistical significance can be obtained by

$$\frac{n_x + n_y - p - 1}{(n_x + n_y - 2)p} T^2, \tag{7.9}$$

where p is the number of variables in each sample.

7.2.4 Effect Size

Besides testing if there is a statistically significant difference between two samples it is also important to determine the strength of the observed phenomenon i.e., the effect size. An effect size is a descriptive statistic that estimates the magnitude of a relationship without making any statement about whether the apparent relationship in the data reflects a true relationship in the population or not. As such, the effect size complements the p-value of statistical significance.

Different measures for the effect size exist and can be either relative or absolute. Examples of these are the odds ratio and the Pearson's correlation. The Pearson's correlation describes a linear correlation between -1 and 1 for a perfect negative/positive linear relation (0 indicates no linear relation). The odds ratio describes the degree of association between two binary variables and is the fraction of the ratios for the appearance of the phenomenon in the one group divided with the ratio of the phenomenon in the other group.

7.3 Dense Mesh Point Correspondence

Many statistical methods, such as PCA, need point (data) correspondence. For 3D scans (meshes), a dense point correspondence between vertices on the different meshes has to be established to use these methods. In this work we use the work by Paulsen *et al.* [121], where correspondence is found by thin plate

spline warping followed by a closest point projection, where the correspondence field is regularized by a Markov random field framework. In short, the method consists of the following steps:

1. Construct a model mesh that will be the basis of the point correspondence.
2. Annotate anatomical landmarks on both the model mesh and the mesh we want to obtain point correspondence to (the target mesh).
3. Warp the model mesh to the target mesh by a volumetric thin plate spline warp using the corresponding anatomical landmarks.
4. Project the vertices from the model meshes onto the target model creating an initial correspondence field.
5. The correspondence field is then regularized by a Markov random field framework and matching curvature.

The regularization ensures that the resulting point correspondence mesh has a minimum of folds and intersections. For a more in-depth description of this method consult [121].

7.4 Human Understandable Facial Distances

From the 74-point annotation scheme described in Appendix F thirty-four human understandable facial distances were constructed. The distances are shown in Figure 7.3 and capture facial distances such as "width of the mouth", "length of the nose", etc. These human understandable facial distances are a good starting point for facial morphology analysis as they are simple and can be clearly understood by doctors and clinical personnel. Both these and some advanced facial morphology measures, such as the point correspondence described in Section 7.3, will be used throughout the studies in this thesis.

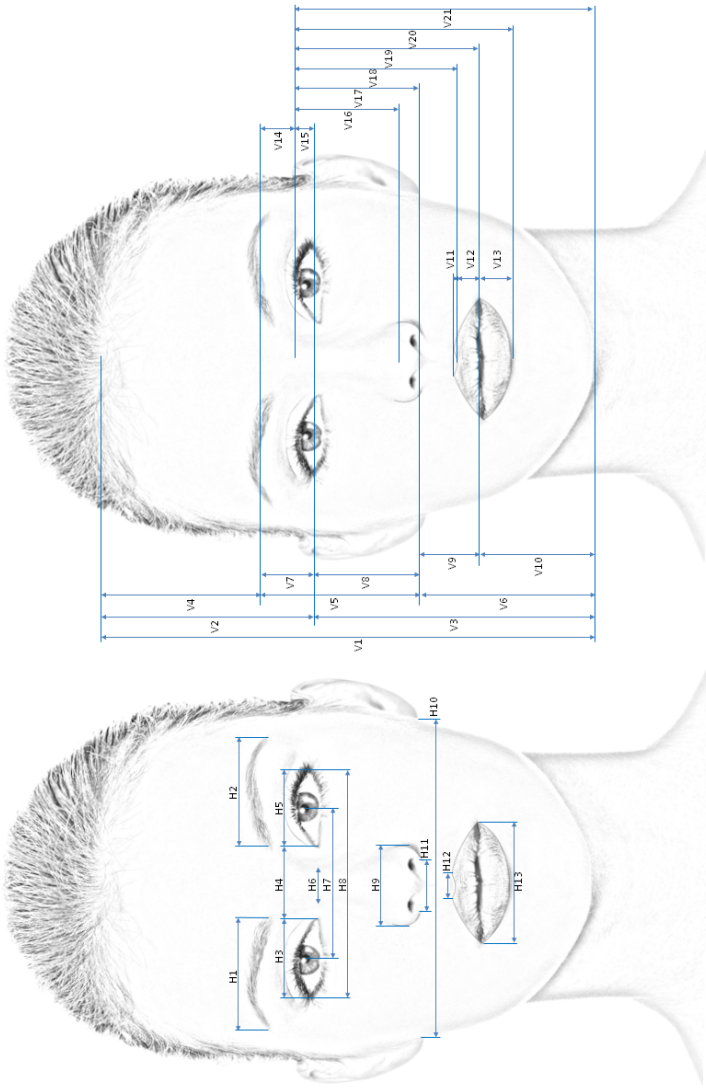


Figure 7.3: Human understandable facial distances.

Part III

Implementation

Chapter 8

Implementation

This thesis includes work from several projects that had to collect and analyze large amounts of 2D and 3D facial data, ranging between 250 and 1500 samples. Software packages and tools were implemented to assist in this process and in some aspects ease the human workload in the data gathering process. A quick overview of these tools is:

AAMLab	a Matlab implementation of two different statistical appearance model frameworks.
FaceDetect	a C++ batch application implementing the Viola-Jones Rapid Object Detection.
AnnoTool2D	a simple light weight annotation application for 2D images.
AnnoTool3D	a simple light weight annotation application for 3D surface scans.

These tools were then integrated into the following setups:

- 2D setup at deCODE Genetics.
- 2D setup at Center for Biological Sequence Analysis at DTU (CBS).
- 3D setups at the Blood Bank and four different Psychiatric Institutions.

As part of this work the author provided consultancy on the design of the 2D setup at CBS and the 3D setups. All software has been developed and tested with respect to stability to a level where they could be included in production at various sites. At the time of writing the first and last setups remain in production at the indicted sites in Iceland and Denmark. In the following, these software tools and setups will be described in more detail.

8.1 AAMLab

AAMLab is a Matlab implementation of two different statistical appearance model frameworks:

- The traditional active appearance models (AAM) algorithm originally proposed by Cootes *et al.* [33].
- A novel fusion of the AAM and the Riemannian elasticity framework, which yields a non-linear shape model and a linear texture model – the active elastic appearance model (EAM).

The novel EAM is documented in Appendix H.

The AAMLab framework was developed by Mads F. Hansen and the author, and is implemented via Matlabs object-oriented programming capabilities for easier use and modification. Some calculation-intensive functions have been implemented in C++ using MEX-files for computational speedup.

The AAMLab framework is now in use at the two data collection setups described in Sections 8.3 and 8.5. Furthermore, the framework is also used in the course "Advanced Image Analysis" (course number 02506) at the Technical University of Denmark. An example of the use of AAMLab to register facial landmarks is shown in Figure 8.1. The AMMLab also includes an explorer to visualize variation in each component in an interactive fashion. An example is presented in Figure 8.2.

Key features of AMMLab include:

- Use linear and/or non-linear algorithms to represent the statistical appearance models (AAM or EAM models).
- Define constrained edges such that models can be constructed within the convex hull of the associated landmark annotations.
- Construct AMM models with shape and texture in combined or separate

principal components.

- Define the percentage of variations the model should capture.

Figure 8.1: Example of the AAMLab framework on human face. Progress left-to-right from initialized mean shape and texture with poor face detection to converged result (the original image is shown to the right).

8.2 2D and 3D Annotation Tools

The manual and automatic annotation tools described in this section are used in many different projects, thus the implementation was done in a flexible manner. In general, they have been constructed for an easy and fast annotation flow with low human processing time. Other existing annotation tools were investigated, such as the landmark functionality in MeshLab[1]. MeshLab is an open source system for processing and editing of unstructured 3D triangular meshes. However, none of the existing software is really designed for use in a production environment where large samples are manually annotated with landmarks.

8.2.1 FaceDetect

FaceDetect is a C++ batch application implementing the Viola-Jones Rapid Object Detection [157, 158]. The implementation uses OpenCV[2], which is an open source computer vision and machine learning software library, and can automatically locate faces in 2D images. When it locates a face it also searches for the following face components: eyes, nose and mouth. Examples of face detection in low resolution images are demonstrated in Figure 8.3. Support for rotated faces is not implemented as it is not required by the work performed in this thesis.

[1] `http://meshlab.sourceforge.net/`.
[2] `http://opencv.org/`.

Figure 8.2: Example of the model explorer in the AAMLab framework. The three images show PC4 of the model being adjusted from the mean (center) to $\mp$ 2 standard deviations.

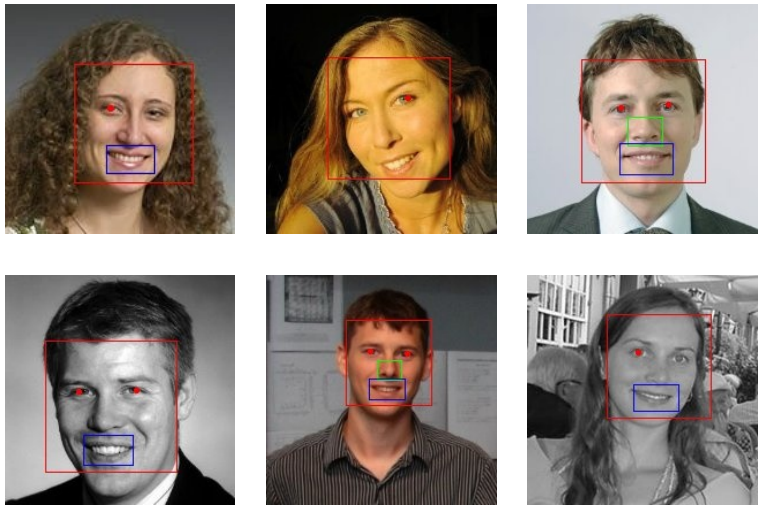


Figure 8.3: FaceDetect batch application used for face and face component detection on low resolution images.

8.2.2 AnnoTool2D

AnnoTool2D is a simple light weight annotation application. It is written in Matlab and compiled into a stand alone batch application. It is created to give an efficient annotation experience for an operator. Existing annotation tools require a series of actions to move annotation points, such as: select point, drag point to new location, release point. This might not seem to be a big issue, but when annotating hundreds of images this workflow becomes tedious. AnnoTool2D makes landmarks adjustment into a "one-click" operation with no need to select landmarks to move them. Examples of the AnnoTool2D are given in Figure 8.4 and Figure 8.5.

Key features of AnnoTool2D include:

- Delaunay triangulation [127] validation. If the Delaunay triangulation changes when one or more landmarks are moved, it and the surrounding landmarks will be highlighted, as the landmark is probably ill-positioned. An example can be seen in Figure 8.6.
- Reposition landmarks with a "one-click" operation.
- Move multiple points on the mouth and nose.
- Jaw, mouth, eye and eyebrow pseudo-landmarks can be moved with con-

necting anatomical landmarks.

- Annotate the pupil and iris as circles, by three landmarks. A circle is uniquely defined by three landmarks that are not collinear.
- The tool can collect meta data by guiding the operator trough questions such as: “Is the face expression neutral (is the person relaxed)?”.

8.2.3 AnnoTool3D

AnnoTool3D is tailored to annotating human faces (spherical objects with a symmetry line), whereas the AnnoTool2D can be used for any 2D annotation. The reason is that we found that landmarks in high curvature areas such as the nose and jaw need to be annotated in 3D, but annotation points in low curvature areas such as mouth, eye and eyebrow can be annotated in 2D and then transformed into 3D with high precision using barycentric coordinates [38] of the facet (triangle) in 3D and the corresponding triangle in the 2D image (texture map). Thus, in the AnnoTool3D only the jaw and nose landmarks are corrected. An example of the AnnoTool3D is shown in Figure 8.7. AnnoTool3D is written in Matlab and compiled into a stand alone batch application that saves the annotation in a Matlab structure format.

Key features of AnnoTool3D include:

- Delaunay triangulation validation, similar to the one in AnnoTool2D.
- Reposition landmarks with a “one-click” operation.
- Enable/disable points, as some scans will have holes that make some landmark “undefined”.
- Different coloring of landmarks left, right and on the symmetry line.

8.3 2D setup at deCODE Genetics

The collaboration between deCODE and the author is focused on associations between the human genome and frontal face images of a cohort of about 1300 Icelandic individuals.

deCODE Genetics is a global leader in genetics research. By their own words they are

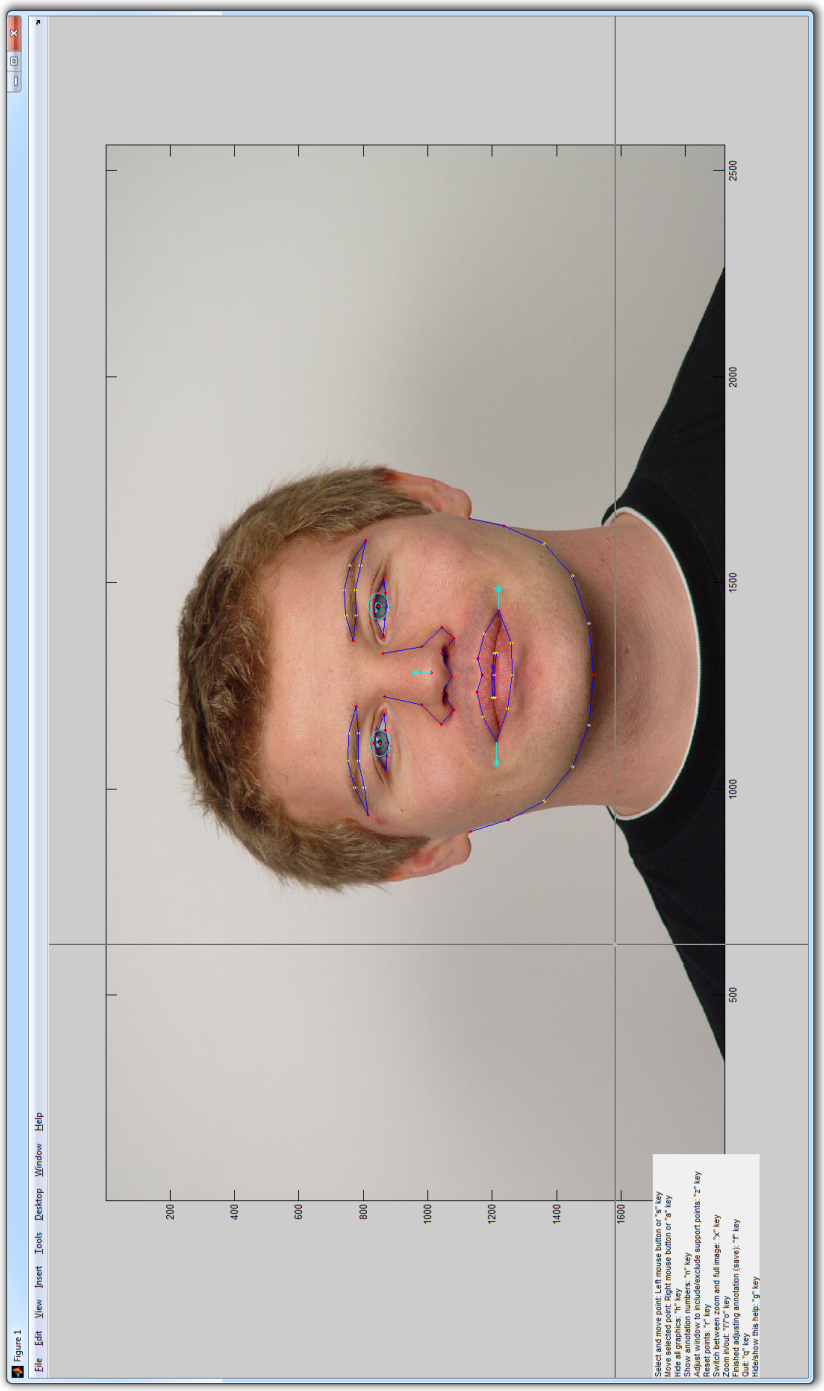


Figure 8.4: Main program window of AnnoTool2D.

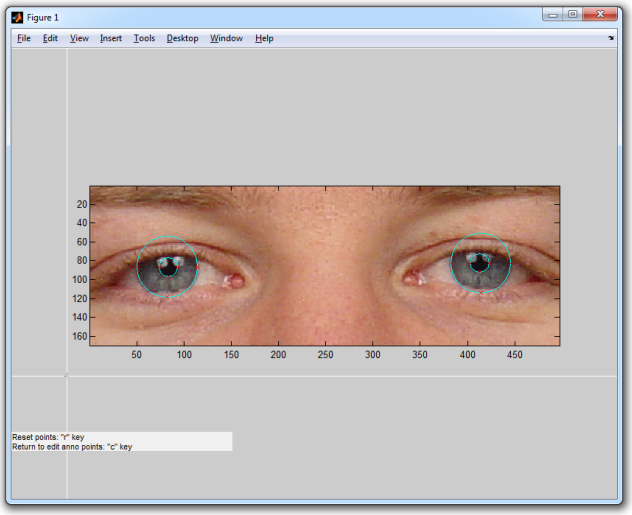


Figure 8.5: Program window of AnnoTool2D used to adjust pupil and iris.

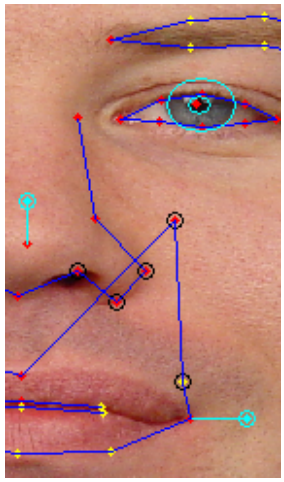


Figure 8.6: Delaunay triangulation validation. If the Delaunay triangulation changes when one or more landmarks are moved, it and the surrounding landmarks will be highlighted by circles, as the landmark is probably ill-positioned.

Figure 8.7: The main program window of AnnoTool3D from a frontal (top) and profile (bottom) view point. The enable/disable point window can be seen on the bottom picture.

> *"... a global leader in analyzing and understanding the human genome. Using its unique expertise and population resources, deCODE has discovered genetic risk factors for dozens of common diseases. The purpose of understanding the genetics of disease is to use that information to create new means of diagnosing, treating and preventing disease."*

In their research they collect DNA samples and many different phenotype categories, such as height, weight, disease diagnose (if any), etc. deCODE has strict policies for anonymity, so employees taking blood samples and interacting with participants do not work with DNA data.

In this work we correlated facial images and distances to DNA. Therefore, we were also put under the strict rules of anonymity, requiring among other things that no data can leave Iceland. Images and DNA data were not allowed to be stored and worked on at the same location. To overcome this working restriction it was necessary to develop a full software suite that at the time of writing remains in production in Iceland. In the following, this software suite and the manual process associated to running the software is described. Steps 1) and 6) are manual. Two Windows batch files where created to run the automatic steps.

1. A frontal picture is recorded with a Canon IXUS 95, 10 megapixels camera, mounted on a custom build rig in a photo studio, where no zoom was used. All participants are recorded sitting on a height adjustable chair, with a horizontal and vertical ruler in the image frame to determine scale.
2. Using camera calibration [80], the image is corrected for radial and tangential distortions introduced by the camera lens. Examples of images from the photo studio before and after correction is shown in Figure 8.8.
3. The eye positions are automatically located by the FaceDetect software described in Section 8.2.1.
4. 74 facial landmarks are automatically found by the AMMLab software described in Section 8.1.
5. The vertical and horizontal ruler are detected automatically by localizing the maximum response in the two-dimensional convolution [83] between the image and a kernel. The kernel as a small cropped image of the ruler's numbers "10" and "20".
6. The facial landmarks are corrected by a human operator (the psychologists employed by deCODE to collect phenotypes) by the AnnoTool2D software described in Section 8.2.2. In the rare cases ($<5\%$) where the ruler is not

automatically located the operator also corrects these points.

7. A batch script prepares and packages facial distances and PCA scores into a transfer file ready to be sent to the DNA site for analysis. The PCA model is not transferred so facial reconstruction is not possible at the DNA site.

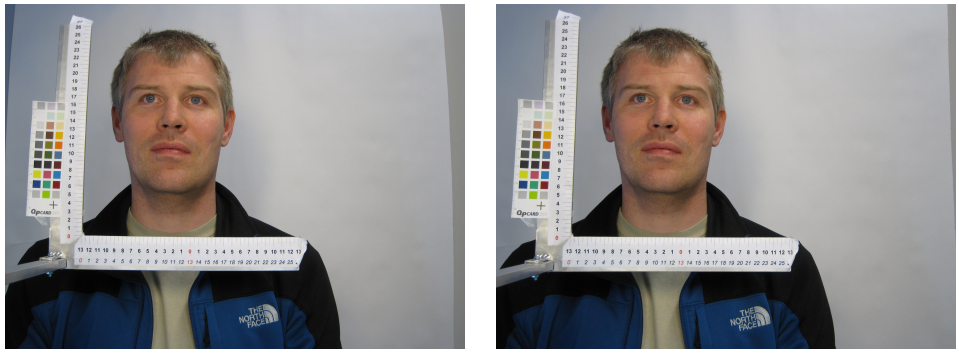


Figure 8.8: Test images from the deCODE photo studio before (left) and after (right) correction for lens distortions.

8.4 2D setup at Center for Biological Sequence Analysis at DTU

A small 2D photo setup was created at the Center for Biological Sequence Analysis (CBS) at DTU. The goal was to create a data set of 244 students at CBS for the use in the following two projects:

- Can you judge a book by its cover? - An investigation into the proposed connection between looks and personality [163].
- Identification of the optimal method for analysis of human faces in the context of finding associations to personality traits [105].

Besides co-supervising these projects the author also provided software for annotation and analysis, similar to the setup on deCODE:

1. A frontal picture is recorded with a Canon PowerShot SX200 IS, 12 megapixels camera, mounted on a tripod in a primitive photo studio, where no

zoom was used. All participants are recorded sitting on a height adjustable chair with a green 8mm circular dot glued to the forehead to determine scale.

2. The eye positions are automatically located by the FaceDetect software described in Section 8.2.1.
3. 73 facial landmarks are automatically found by the AMMLab software described in Section 8.1.
4. The facial landmarks are corrected by a human operator by the AnnoTool2D software described in Section 8.2.2. Wiktor Majewski performed all the annotations in connection with the work done in [105].
5. AMMLab software was then used to construct appearance models that in turn were used in the two projects.

No further description of these projects will be presented in this work.

8.5 3D setups at the Blood Bank and Psychiatric Institutions

As part of the work with the Institute for Biological Psychiatry (iBP), Mental Health Center Sct Hans, Copenhagen University Hospital a semi-portable setup for data collection was developed. The setup was able to collect the following data:

- 3D facial scans using a Canfield Vectra M3 Imaging System.
- 2D facial and hand images using a Nikon D90.
- Anthropometric data using laser rulers and measuring tape.

To collect the data in a seamless way several small tools were developed by the author:

2D Image Capture is a small remote control application for the Nikon D90. The application allows the operator to capture an image via the Nikon D90, then transfer it to the computer and display the image in the application. If the image is not good, a new image can be captured that will overwrite the existing image on the computer.

Process 2D & 3D Images is a small image data management tool. It collects the participants unique ID via a handhold bar code reader. It moves and renames the 2D image captured by `2D Image Capture` and the "3D scan" files to a folder with the participants unique ID.

Antropometrisk Data is Danish for *Anthropometric Data* and collects the participants height, wingspan and head circumference. This application is an adaptation of the `Process 2D & 3D Images` application and is made by Thomas F. Hansen, iBP.

All the tools were developed in HTML and javascript encapsulated in an HTML application. Screenshots of the applications are shown in Figure 8.9. An image of the Canfield Vectra M3 Imaging System can be seen in Figure 8.10.

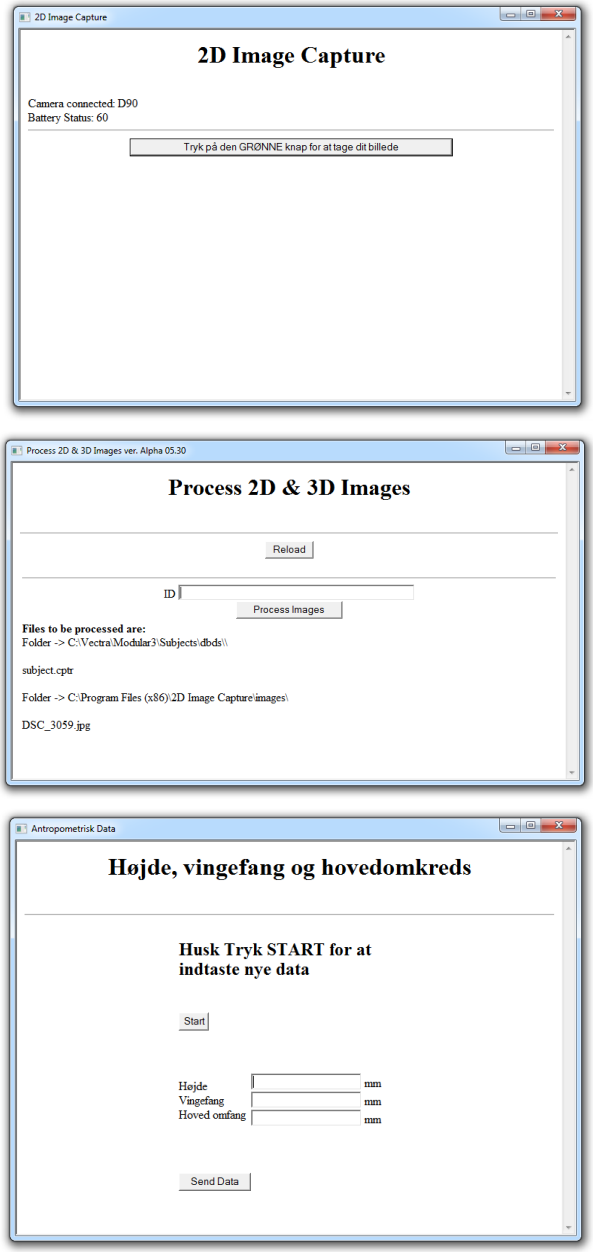


Figure 8.9: Small HTML applications developed to collect images, 3D scans and anthropometric data.

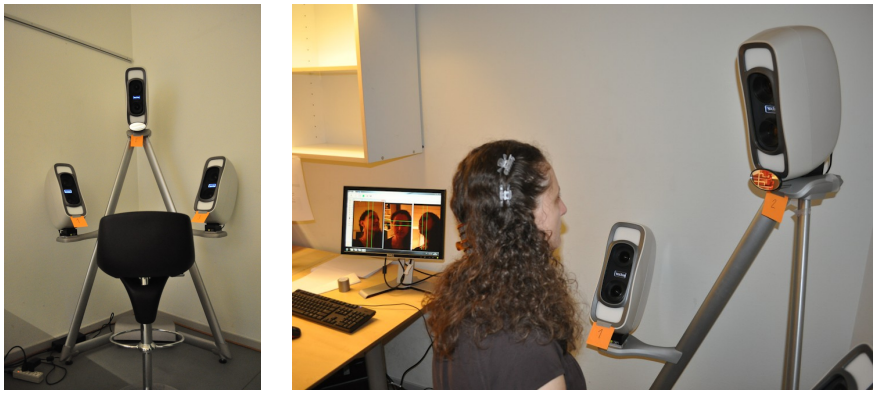


Figure 8.10: The Canfield Vectra M3 Imaging System.

Part IV

Discussion and Conclusion

Chapter 9

Discussion and Conclusion

This thesis has addressed a few of many missing links in *Human Biometics* and *Human Genetics* with regards to facial analysis. This final chapter discusses the presented results and ends with concluding remarks.

9.1 Discussion

In this work multiple subjects have been addressed and studied. We take a "non-holistic" approach to facial recognition by sparse PCA, which produces a more human understandable features sub-space, and we show that it can obtain the same performance as the dense PCA version, using only a fraction of the input data. By trying to understand the visual part of the human cognition process of gender recognition we showed the importance of training machine learning algorithms with data that do not contain ill-defined samples (low quality training data). We used human cognition to define a new metric called "gender strength", which we use to refine data sets (remove ill-defined samples). This resulted in significantly higher performance on known machine learning algorithms, which indicates that not only the volume of the training data is important but also the quality of the training data. Furthermore, we determined the human cognition capability to recognize gender in a split second (200 ms) to be quite remarkable. Only about 20 % misclassification was observed even though the test

participants reported that they felt like they were guessing most of the time. Our results show that on high resolution data machine learning algorithms outperform human operators, whereas on low resolution data they have similar performance. This indicates that the machine learning algorithms exploit the available information in the high resolution images more efficiently compared to humans.

Next, the variability of 3D facial landmarks has been examined with respect to intra- and inter-operator variability. It is important to understand the variability embedded in manual annotation, because registration, recognition, and machine learning algorithms can be influenced by manual annotation errors. A proof-of-concept study on the correlation of complex polygenic variations to specific holistic facial characteristics was undertaken. It showed that to some extent we could predict facial characteristics based solely on genetic information. This form of analysis may at some point result in the prediction of portraits resembling an individual's real face, which may be used for screening suspects for solving crimes, but probably never for identification. We then looked at the discrete facial feature of NRNX1 deletions carriers on a small data set and the detection of minor physical anomalies in the face associated with severe schizophrenic disorders on a large data set, both showing promising results, which suggest that facial characteristics can be associated with these phenotypes. The obtained high performance of 87.4% correct classification suggests that looking for minor physical anomalies in high accuracy 3D scans can give a risk indication that could assist doctors in diagnosing schizophrenic disorders.

Finally, we did provide the research community with a Matlab implementation of AAM and EAM, a linear and a non-linear appearance models.

The existing work in *Human Biometics* is much more comprehensive than the work in *Human Genetics* with regards to facial analysis. Therefore, we took a different human cognitive approach in the work related to biometics, deriving interesting results. The work with understanding genetic variations with regards to facial analysis has been very pioneering, with proof-of-concept study and extraordinary findings. However, these results have been created on small or medium sized data sets that hopefully will be expanded in the future, yielding more and stronger findings.

9.2 Conclusion

The first objective of this thesis was to progress our understanding of biometric recognition based on facial data from a human point of view, which has been

addressed in three contributions given in Appendix A, B and C. An overview of these contributions together with an analysis of particular advantages and/or disadvantages can be found in Chapter 2.

The second and third objectives of this thesis were to progress our understanding of the human genome with respect to facial morphology and to asses the potential of detecting severe schizophrenic disorders from minor physical anomalies in the face. These have been addressed in three contributions in Appendix E, F and G and an overview of these contributions together with an analysis of particular advantages and/or disadvantages can be found in Chapter 2.

The last objective of this thesis was to construct data sets containing both high resolution 3D scans and genotype data. The data sets constructed during this thesis consist of approximately 1000 control subjects and approximately 215 patients with schizophrenic disorders and are used in the work in conjunction with the second and third objectives. The software packages and hardware setups enabling the creation of the data sets are described in Chapter 8. Furthermore, two contributions on the variability of human annotation (see Appendix D) and an AAM/EAM implementation (see Appendix H) were done to support the construction of the data sets.

As commercial 3D scanners with high resolution and accuracy become more accessible in price and quality, it is our belief that facial analysis will become a valuable tool assisting doctors and physicians in their study, diagnosis, treatment, and maybe even prevention of mental disorders. This is not a reality today, but in the light of the obtained results (high classification capability of about 87%) a currently unused potential can be seen.

With the completion of this thesis we do not close the book on the addressed issues but rather indicate the enormous potential we one day will be able to yield within these fields.

We believe that our findings provide an important stepping stone in the further exploration of the links between different phenotypes and our genetic markup.

Part V

Appendices

Appendix A

Sparse Similarity-based Fisherfaces

Jens Fagertun, David D. Gomez, Mads F. Hansen and Rasmus R. Paulsen

Abstract

In this work, the effect of introducing Sparse Principal Component Analysis within the Similarity-based Fisherfaces algorithm is examined. The technique aims at mimicking the human ability to discriminate faces by projecting the faces in a highly discriminative and easy interpretative way. Pixel intensities are used by Sparse Principal Component Analysis and Fisher Linear Discriminant Analysis to assign a one dimensional subspace projection to each person belonging to a reference data set. Experimental results performed in the AR dataset show that Similarity-based Fisherfaces in a sparse version can obtain the same recognition results as the technique in a dense version using only a fraction of the input data. Furthermore, the presented results suggest that using SPCA in the technique offers robustness to occlusions. Keywords: *Face recognition, Sparse Principal Component Analysis, Fisher Linear Discriminant Analysis, Biometrics, Multi- Subspace Method.*

A.1 Introduction

Recognizing a face is an important everyday task for human interaction. It is a skill we acquire before we can walk, and we are able to perform it with high

accuracy using little or no effort. Due to the importance of this skill, machine aided face recognition is one of the most researched fields in image analysis. However, results reported in literature suggest that facial recognition still lacks the performance that a human operators can achieve.

In the last two decades, as advanced spectral techniques have emerged, there have been a gradual shift of the research on face recognition from geometrical towards spectral analysis. Such methods include the unsupervised method Eigenfaces [95, 151] and the supervised method Fisherfaces [12]. Later on, related techniques have been proposed aiming at obtaining better classification results by developing more discriminative projections [27, 103]. Most supervised methods in the literature today try to project the face representation into a subspace where a measure of global separation is maximized. However, the proposed algorithm in this paper tries to project the face representation into a series of one-dimensional subspaces where one person (class) is discriminated from all others in the population.

By solving the problem in a series of subspaces, this algorithm makes the enrollment or removal of a person easy. It is not needed to recalculate all of the existing individual subspaces when changes to the population are made. When a new person is introduced, simply a new individual subspace is added. If a person has to be removed from the database, it is only needed to remove the corresponding individual subspace.

In this article the algorithm of Similarity-based Fisherfaces proposed in [68] is extended by introducing Sparse Principal Component Analysis (SPCA) into the recognition algorithm. The terms *dense* and *sparse* will be used to differentiate between standard Principal Component Analysis (PCA) and SPCA.

The structure of the paper is as follows. Section A.2 describes the algorithm to construct the sparse similarity-based face representation. Section A.3 presents results that show the discriminative power of using SPCA versus PCA in Sparse Similarity-based Fisherfaces and its ability to discover the individuals most discriminative characteristics. Section A.4 gives a discussion and conclusion for Sparse Similarity-based Fisherfaces.

A.2 Algorithm Description

The proposed algorithm builds upon SPCA [90] and Fisher Linear Discriminant Analysis (FLDA) [12].

However, unlike traditional FLDA, which maximizes a global measure of class separation[1], it obtains a one-dimensional individual linear subspace for each person (class) enrolled in a training database. The value of the projection of each sample (face representation) projected into this subspace aims at measuring the similarity of the sample with respect to that person for whom the subspace was created. Following, the proposed algorithm is briefly described. For a better understanding, Fig. A.1 displays a diagram of the algorithm.

A.2.1 Obtaining the Texture Formulation

Face Recognition can be conducted using several types of features such as geometrical and textural[2]. In this work, visible texture features are used. The texture features are obtained by a piece-wise affine warp based on the Delaunay triangulation of the mean shape. Hereafter, the texture is normalized to zero mean and unit variance.

When the facial feature representations has been obtained, these are projected into a SPCA feature space to remove redundancy. There are different ways to formulate the SPCA objective function [171, 90]. In this study the generalized power method for SPCA formulation is used as described in [90]

$$\phi_{\ell_0}(\gamma) = \max_{x \in S^P} \sum_{i=1}^{n} [(a_i^T x)^2 - \gamma]_+. \tag{A.1}$$

Formulation (A.1) yields a single Sparse Principal Component (SPC). In order to obtain more than one SPC, (A.1) is used iteratively by means of deflation. Deflation is obtained as follows: when a new SPC is obtained its variation is removed from the populations variance before obtaining the next SPC. The advantage of this is that the resulting SPC number k will be nearly orthogonal to the $k-1$ previous computed SPC.

The γ parameter controls the sparsity of the solution in such a way that the size of γ will move the SPCA solution between the two extremes: a full PCA and the zero solution. It depends on the application whether sparsity or explained variance is valued.

[1] In a subspace of dimensionality "Number of classes" minus one.
[2] Textural features can be recovered from any spectral range (E.g. visible or infrared).

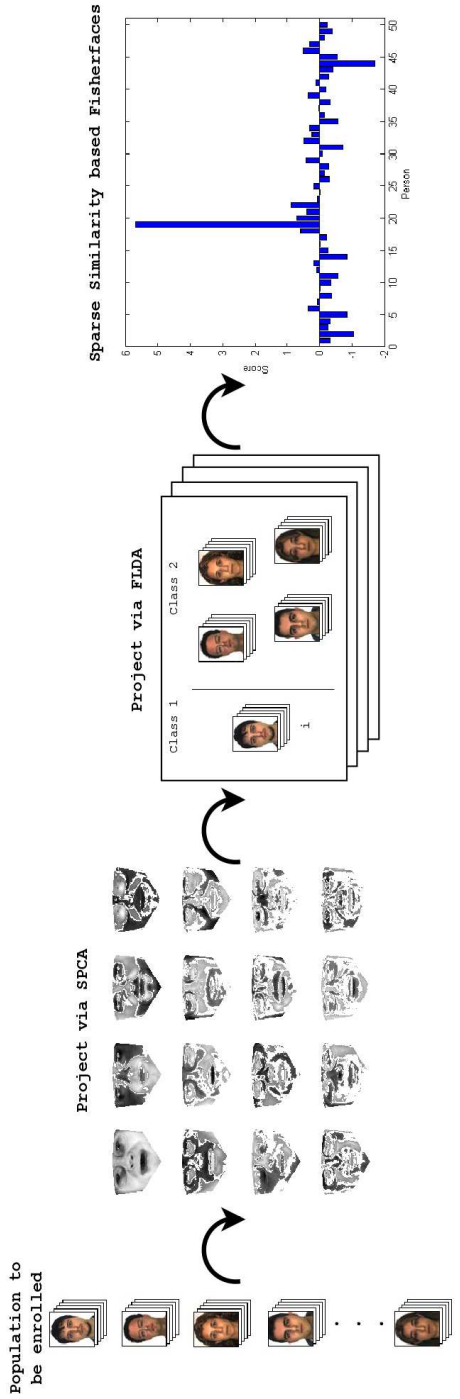


Figure A.1: Algorithm overview. SPCA and FLDA is used in turn on the population to be enrolled, in order to obtain the Sparse Similarity-based Fisherface representations.

A.2.2 Creating the Individual Subspaces

After the face representations have been projected into the SPCA feature space, a second projection using FLDA for each individual in the database is conducted to build a personalized subspace for each of them. This subspaces are obtained discriminating each individual with respect to all others.

A standard FLDA projects the data samples into an F-dimensional subspace so that, it maximizes the ratio of the between-class scatter to the within-class scatter. The dimensionality F of this subspace is equal to the minimum of n - 1 and m - 1, where n is the number of variables and m is the total number of people (classes). The projection matrix W is found by maximizing the ratio

$$\frac{W^T S_B W}{W^T S_W W}, \tag{A.2}$$

where S_B and S_W are the between-class scatter and the within class scatter matrices, respectively. The projection vectors of this matrix W correspond to the eigenvectors associated to the non-zero eigenvalues of the matrix $S_W^{-1} S_B$.

In this case of two classes FLDA returns a one dimensional subspace for each individual. For a more formal discussion of FLDA and Individual Subspaces see [68].

A.2.3 Classification

To turn the obtained projections into measurements of similarity a standardization is applied. The standardization of model $i = 1, \ldots, m$ is based on two assumptions. First, the number of observations for person i is much smaller than the number of the observations of all other people. Second, the projection of the other people follows a Gaussian distribution. These two assumptions imply that the distribution of all the projected facial images on a particular discriminative individual model is a Gaussian distribution with outliers. The standardization of model i is then achieved by transforming the projections into a standard Gaussian distribution, keeping the projections of the person i positive. Formally, let $\bar{x}_i$ be the mean of the projections on model i, σ_i the standard deviation, and let $x_{i,j}$ be the projection of face representation j in the ith subspace. These projections are standardized by

$$\hat{x}_{i,j} = (x_{i,j} - \bar{x}_i)/\sigma_i. \tag{A.3}$$

If the standardized projection for the images corresponding to person i are negative, then $\hat{x}_{i,j}$ are replaced by $-\hat{x}_{i,j}$ for all projections. This causes the

projection of the images corresponding to person i to be positive and far from the mean of the gaussian.

Once the model i is standardized, the probability of a projected image to belong to person i is given by the value of the standard normal cumulative function in the projected value. This fact is used to classify a given image. If it is assumed that the image belongs to a person from the data set, the image is projected by all the models and classified as belonging to the model that gives the largest probability. Moreover, it is also statistically possible to decide if a given person belongs to the data set or it is unknown. This can be achieved by comparing the largest projection obtained in all the models with a probabilistic threshold. For example if a 99.9% of probability is required, a given image will only be considered as belonging to the database if the projection in one of the individual models is higher than 3.1 standard deviations.

A.3 Experimental Results

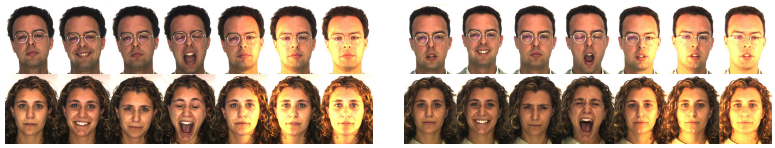
In this article two experiments are presented. The first experiment aims at determining classification rates of sparse vs. dense versions of Eigenfaces, Fisherfaces and Similarity-based Fisherfaces, together with a visualization of Sparse Principal Components. The second experiment analyzes and visualizes which are the most discriminating pixels in a face image based on the algorithm for Sparse and Dense Similarity-based Fisherfaces.

A.3.1 Classification Accuracy

This experiment aims at comparing the performance of SPCA and PCA in the proposed method with respect to Fisherfaces and Eigenfaces method in terms of false classification rates. Both PCA and SPCA versions of Fisherfaces and Eigenfaces will be used in this experiment.

As data set for this study 50 persons (25 male and 25 female) was randomly selected from the AR face database [108]. The database is composed of two independent sessions recorded 14 days apart (Only images without occlusions is used). An example of the selected images for two persons is displayed in Fig. A.2. All the images were manually annotated with 22 landmarks.

The data set was divided into two sets. The images of the first session were



(A) (B)

Figure A.2: The AR data set: (A) The seven images without occlusions from first session, (B) The seven images without occlusions from the second session.

used to train the algorithms, whereas the images from the second session were subsequently used to test the performance. In order to obtain the texture representation of each face in the training set, the different images were warped with respect to the mean shape, represented by 41339 pixels. These representations was normalized to zero mean and unit variance.

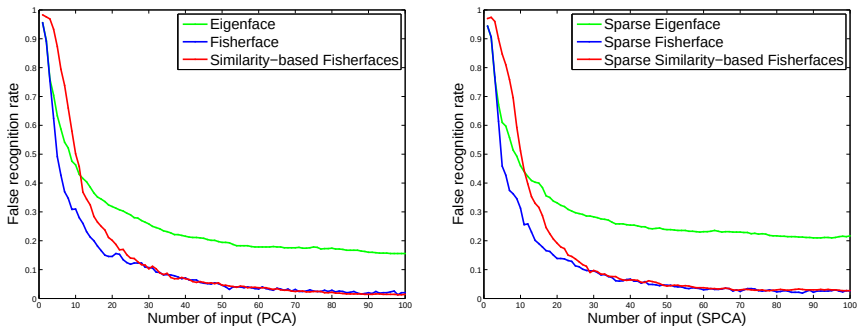
In the Fisherface and the Eigenface algorithms the Nearest-Neighbor algorithm with Euclidean metric was used as classifier. In the proposed method the classification is performed in such a way; that a given face image is recognized as the person associated to the subspace that yields the highest probability.

The test was repeated a second time changing the roles of the training and the test sets: session two was used as training data and session one as test data.

A.3.1.1 Determine γ for Use in SPCA

Considering the formulation for SPCA (A.1) it can be determined that as γ goes to zero the SPCA goes to a full PCA solution. As described before, the process of choosing γ is not trivial and depends on the problem at hand, whether a more dense or sparse solution to SPCA is wanted. After performing exploratory experiments with different values of γ the value of $\gamma = 0.003$ was chosen as a good value for this problem. It is beyond the scope of this paper to analyze methods for determining an optimal γ for a given problem. The value of $\gamma = 0.003$ is used for the experiments in this study.

The false classification rates for the different techniques are shown in Fig. A.3, where these rates are plotted as a function of using the first i PC/SPC. The figure shows rates up to the first one hundred PC/SPC.



(A) (B)

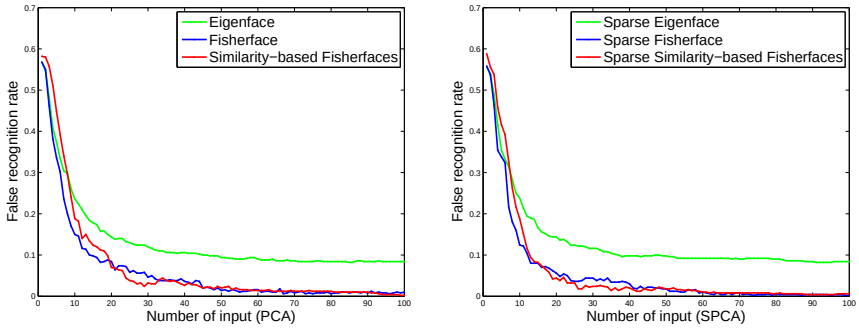
Figure A.3: False classification rates for Eigenfaces, Fisherfaces and Similarity-based Fisherfaces with extreme light variation. (A) displays the Dense versions, (B) displays the Sparse versions. The rates are plotted as a function of using the first i PC/SPC, $i = 1 \ldots 100$.

To analyze the effect of the light variance, another test was conducted where the three pictures containing extreme light of each person in the test and training set was removed. This reduces the inner class variance, due to lighting noise. Any method for postprocessing could be used for removing light variation. However, for simplicity the images are simply removed in this study. The new false classification rates for the different techniques are shown in Fig. A.4. The SPC can be seen in Fig. A.5 for the two experiments with and without extreme light, respectively.

From Fig. A.3 it can be seen that the Sparse versions of Fisherfaces and Similarity-based Fisherfaces obtain similar recognition results as the Dense versions with only using a fraction of the input data. In Fig. A.4 it can be seen that Sparse Similarity-based Fisherfaces preforms even better than the Dense version when large variance in the sample population due to inner class variance noise is minimized (removing light variation). The false classification rates for the first 25, 50 and 100 PC/SPC in Table A.1 and Table A.2 for the different techniques with and without extreme lighting, respectively.

A.3.2 Discriminative Pixels

An interesting property of the proposed algorithm is that it is possible to determine which are the most discriminative features of a given person. The 10,



(A) (B)

Figure A.4: False classification rates for Eigenfaces, Fisherfaces and Similarity-based Fisherfaces without extreme light variation. (A) displays the Dense versions, (B) displays the Sparse versions. The rates are plotted as a function of using the first i PC/SPC, $i = 1 \ldots 100$.

Method	**Dense version**			**Sparse version**		
	25 PC	50 PC	100 PC	25 SPC	50 SPC	100 SPC
Eigenfaces	29%	19,4%	15,6%	29,7%	23,9%	21,7%
Fisherfaces	11,9%	4,6%	2%	11,3%	4,6%	2,6%
Similarity-based Fisherfaces	14%	4,6%	1,4%	13,7%	4,3%	2,7%

Table A.1: False classification rates for Eigenfaces, Fisherfaces and Similarity-based Fisherfaces with extreme light variation. The rates are shown for the first 25, 50 and 100 PC/SPC, respectively.

Method	**Dense version**			**Sparse version**		
	25 PC	50 PC	100 PC	25 SPC	50 SPC	100 SPC
Eigenfaces	13%	9,4%	8,4%	12,4%	9,8%	8,4%
Fisherfaces	5,8%	1,4%	1%	3,6%	1,8%	0,2%
Similarity-based Fisherfaces	3,8%	2%	0,2%	3,2%	1,8%	0,6%

Table A.2: False classification rates for Eigenfaces, Fisherfaces and Similarity-based Fisherfaces without extreme light variation. The rates are shown for the first 25, 50 and 100 PC/SPC, respectively.

15 and 25% discriminative pixels corresponding to the highest weights in the model are displayed (in red) in Fig. A.6 for Sparse and Dense Similarity-based Fisherfaces respectively. It is clear that important discriminating features include eyes, noses, glasses, moles and beards. Notice that the Dense algorithm detects the glasses and the mole of person 43 in Fig. A.6 as discriminative features, whereas the Sparse algorithm does not do this to the same degree. This suggests that the Sparse method will be more robust giving the same result no matter if the person is wearing glasses in one picture and contact lenses in another.

A.4 Discussion and Conclusion

In this work, the effect of introducing Sparse Principal Component Analysis within the Similarity-based Fisherfaces algorithm are examined. Both the dense and sparse version of Similarity-based Fisherfaces aims at being a precise and robust algorithm that can be incorporated into biometrical security systems.

Experimental results in this study have shown that the technique in a sparse version can obtain the same recognition results as the technique in a dense version (presented in [68]) with only a fraction of the input data.

Furthermore, the presented results suggest that using SPCA in the technique offers robustness to occlusions. The discriminative pixels are not primarily fixed on naturally occurring occlusions in the face (Glasses, molds etc.), whereas for the dense version of the algorithm they are. This point needs to be examined in detail in future work.

Moreover, just as the dense version, the sparse version also allows for a simple interpretation of the results in the final one-dimensional individual subspace.

Another interesting property of the proposed algorithm is that by solving the problem in a series of subspaces, this algorithm makes the enrollment or removal of a person easy. It is not needed to recalculate all of the existing individual subspaces when changes to the population are made. When a new person is introduced, simply a new individual subspace is added. If a person has to be removed from the database, it is only needed to remove the corresponding individual subspace.

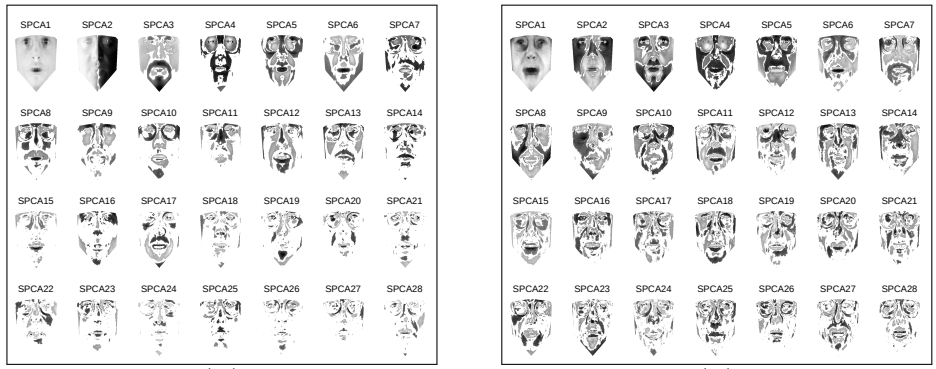


Figure A.5: Visualization of the first 28 Sparse Principal Components from the experiment (A) with extreme light variation (B) without extreme light variation.

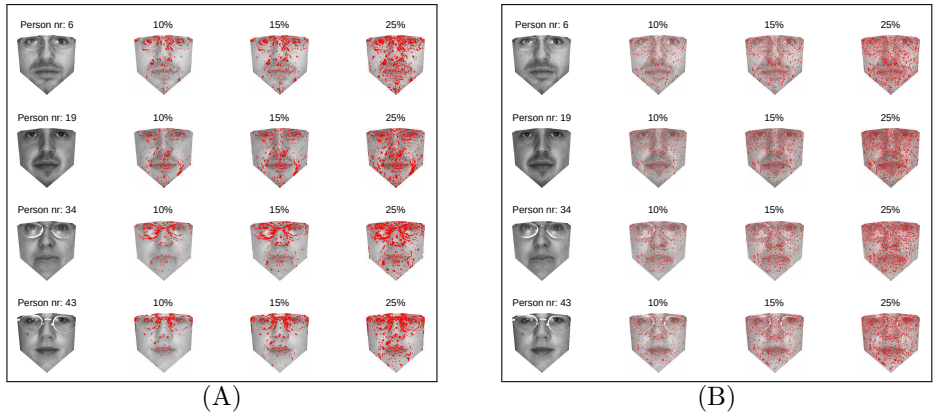


Figure A.6: The 10, 15 and 25% discriminative pixels displayed in red for (A) Similarity-based Fisher and (B) Sparse Similarity-based Fisherfaces.

Appendix B

Gender recognition using cognitive modeling

Jens Fagertun, Tobias Andersen and Rasmus R. Paulsen

Abstract

In this work, we use cognitive modeling to estimate the *"gender strength"* of frontal faces, a continuous class variable, superseding the traditional binary class labeling. To incorporate this continuous variable we suggest a novel linear gender classification algorithm, the Gender Strength Regression. In addition, we use the gender strength to construct a smaller but refined training set, by identifying and removing ill-defined training examples. We use this refined training set to improve the performance of known classification algorithms. Also the human performance of known data sets is reported, and surprisingly it seems to be quite a hard task for humans. Finally our results are reproduced on a data set of above 40,000 public Danish LinkedIN profile pictures. Keywords: *Gender recognition, Linear Discriminant Analysis, Support Vector Machines, Cognitive Modeling, Linear Regression.*

B.1 Introduction

Determining gender is a natural, subconscious task performed everyday in human interaction. By continuously training of this task on a daily basis, the

human mind is shaped to become a gender recognition expert. Despite the strong belief that this is the case, there are very few scientific studies to support this. The studies that make a comparison between human vs. machine gender recognition performance are nearly two decades old, such as [20]. However, it is probably safe to assume that we humans do perform well as gender recognition experts, even with limited scientific documentation of the subject. This gives rise to some interesting questions: "Can we understand how humans perform the task of gender recognition?" and "Can we use this information to enhance machine learning models?".

It should be noted, that when confined only to non-intrusive data (such as visual or audio data) the two populations (male and female) cannot be perfectly separated. There are border-line cases where determining gender is also extremely hard for humans; cases where sports athletes need to undergo a DNA test to exactly determine the gender, come to mind.

The problem of automatic gender classification has been studied for years. Many algorithms have been proposed that employ different methods, including Linear Discriminant Analysis (LDA) [11], Haar-like wavelets [106], Locally Binary Patterns (LBP) [106] and Support Vector Machines (SVM) [114]. All these methods work on a binary understanding of class association, which does not necessarily enable the method to exclude ill-defined samples that could lead to a performance decrease. This work tries to address this shortcoming by analyzing how humans perform gender classification and incorporating this information into both a novel and well-known pattern recognition algorithms.

Obviously, humans perform gender recognition based on a variety of data, such as face, hair style, clothes, height, voice and movement patterns. In this article we look exclusively at faces. To address the previously posed questions, this work will look into the human performance. We estimate a continuous *"gender strength variable"* and use it to construct a novel linear gender classification algorithm, the Gender Strength Regression (GSR). In addition, we look into using this gender strength variable to construct a smaller but refined training set, by identifying and removing ill-defined training examples. We then assess if this refined training set improves performance of known classification algorithms.

The structure of the paper is as follows. Section B.2 introduces the data sets used in this work. Section B.3 describes the cognitive test that estimates the gender strength variable. Section B.4 presents the Gender Strength Regression method. Section B.5 presents the experiments and the obtained results. Finally,

Section B.6 presents a discussion and conclusion on this work.

B.2 Data sets

In this work we use four data sets in order to ensure that our results are not biased towards a single protocol of acquiring facial images. The data sets are subsets of publicly available data, filtered so that they only contain frontal images with no extreme lighting or extreme facial expressions. The images have been cropped and rotated to only show the central face, excluding outer cheeks, lower chin and hair. Every image has been converted to gray scale and standardized by a histogram equalization. Furthermore, the data sets include the same number of males and females. This should motivate the observer to employ a response criterion that is not biased towards one of the two response categories. The data sets used are described in Table B.1 with random samples shown in Figure B.1.

Data set	Total number of images used
AR [108]	298
XM2VTS [112]	552
FERET [124]	836
Public Danish LinkedIN profile pictures	200

Table B.1: Data sets used in this work.

B.3 Cognitive test

The object of the cognitive test is to estimate the gender strength of each sample. The gender strength variable is defined as a variable in the interval [0 1] where close to zero signifies belonging poorly to the sample's group, whereas one signifies belonging strongly to the sample's group. To encode the gender in the variable, females will be denoted in the negative interval [-1 0] and males in the positive interval [0 1].

In order to estimate the gender strength variable, two cognitive user tests were devised and conducted on four and six test participants, respectively. The gender strength variable is estimated as a latent variable encoded by the time it

Figure B.1: Random sample images from the four data sets, with females and males in the first and second row, respectively.

takes participants to perform the gender classification task.

The cognitive test is conducted with the psychtoolbox for matlab [19], to ensure precise exposure/answer times. In the first test (time limited) every facial image is shown for precisely 200 ms, followed by a scrambled image of the original image (also shown for 200 ms) to disrupt the after image, which is sustained neural activity in the visual system following a visual stimulus. In the second test (not time limited) every facial image is shown until the user makes the classification. A graphical illustration of the tests is shown in Figure B.2.

The gender strength variable is estimated by standardizing the test persons´ answer times, so that all answer times from one test person (per data set) has mean zero and unit variance. Then, a pooling of the standardized answer times from all test persons is performed. Finally, the pooled answer times below the mean are encoded to one, the outliers (above three standard deviations) are encoded to zero and the remaining are linearly encoded between the values zero and one.

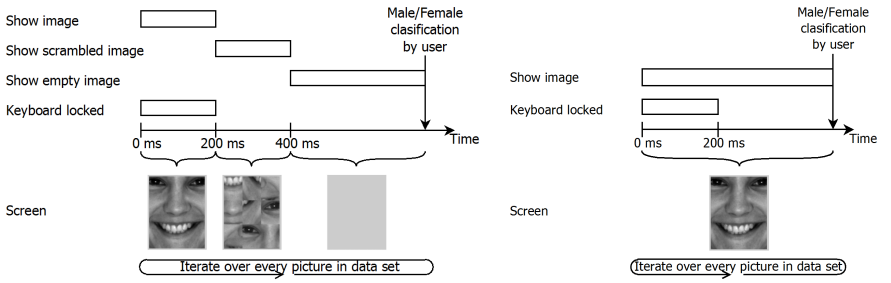


Figure B.2: Timeline for the cognitive tests, (a) time limit and (b) no time limit. There is a 550 ms delay from a user classification to the next picture is shown.

B.4 Gender Strength Regression

In classification problems, training data is usually only labeled with which class it belongs to. As a result, classifiers, such as LDA and SVM, have only a binary understanding of class membership, which they use in the objective function to construct the classification method. LDA seeks to minimize the within class variance while maximizing the between class variance, whereas SVM tries to maximize the margin (distance) between classes. Due to this binary understanding of class, these methods work on the complete data set. The performance of methods such as LDA and SVM depends solely on the given training data, where more training data does not necessarily yield a better performance.

In this paper we suggest enhancing the performance of classifiers such as LDA and SVM by removing ill-defined training samples. In this work, ill-defined samples are defined as cases where more than 50% of the test participants in the the human cognitive tests fail to perform a correct classification.

Furthermore, we suggest a new classifier that seeks to estimate the gender strength variable by regression. A linear least squares regression has the objective function

$$\min_{\beta} \|\beta \mathbf{X} - \mathbf{y}\|^2, \tag{B.1}$$

with the closed form solution

$$\beta = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}, \tag{B.2}$$

where $\mathbf{X}$ is a matrix with samples and variables in the rows and columns, respectively. $\mathbf{y}$ is a vector with the gender strength scores, also called the response variable. β is the coefficient for the regression line. In an optimal system ((B.1) is small) the decision threshold would be zero in accordance with how we encoded the response variable in the regression, see Section B.3. However, this will not always be the case. A more accurate decision threshold can be calculated by

$$\frac{\sum_{i=1}^{k}\beta\mathbf{x}_i}{2k} + \frac{\sum_{j=1}^{l}\beta\mathbf{x}_j}{2l}, \tag{B.3}$$

where i and j belong to samples of class male and female, respectively. k and l are the sizes of the two classes. This assumes that the two classes have the same distributions.

B.5 Experiments

B.5.1 Cognitive test

In the cognitive test, the False Classification Rate (FCR) is reported in Table B.2. It can be seen that in the first test where participants only saw the central face for a fraction of a second the performance was not bad, in the sense that it is clearly better than guessing (test participants reported that they felt like they were guessing most of the time). When the time constraint was removed, the performance increased. Also it seems that the individuals in the AR and Feret data sets had a bias toward masculinity, resulting in more females that were wrongly classified as a male than compared to the two other data sets.

Data set		Test with time constraint	Test without time constraint	Combined
AR	Male*	0.068 ± 0.047	0.006 ± 0.008	0.031 ± 0.042
	Female**	0.107 ± 0.062	0.164 ± 0.070	0.141 ± 0.070
	Total	**0.174 ± 0.096**	**0.169 ± 0.066**	**0.171 ± 0.074**
XM2VTS	Male*	0.084 ± 0.054	0.035 ± 0.015	0.055 ± 0.042
	Female**	0.097 ± 0.045	0.053 ± 0.030	0.071 ± 0.041
	Total	**0.181 ± 0.050**	**0.088 ± 0.025**	**0.126 ± 0.059**
Feret	Male*	0.048 ± 0.052	0.027 ± 0.009	0.035 ± 0.033
	Female**	0.163 ± 0.131	0.079 ± 0.039	0.112 ± 0.092
	Total	**0.211 ± 0.121**	**0.106 ± 0.043**	**0.148 ± 0.094**
Linked-InDK	Male*	0.059 ± 0.038	0.038 ± 0.020	0.046 ± 0.029
	Female**	0.104 ± 0.048	0.028 ± 0.019	0.058 ± 0.050
	Total	**0.163 ± 0.053**	**0.065 ± 0.028**	**0.104 ± 0.063**

* Male wrongly classified as a female.
** Female wrongly classified as a male.

Table B.2: False classification rates for the cognitive test with standard deviations.

B.5.2 Projection lines

To obtain an understanding of how GSR classifies gender, all samples in the data sets are projected onto the respective GSR line. Only the time limited test is used as we want to see the first features we as humans use to perform gender classification. The estimated Gaussian distribution of these projections is shown in Figure B.3, together with synthesized faces corresponding to points on the regression line.

B.5.3 Recognition results

B.5.3.1 Cross-validation evaluation

Gender classification was performed for all data sets using GSR and compared to LDA and SVM[3] with a linear kernel in a leave-one-out cross-validation scheme (Train on all data, except one sample. Then test on this "unseen" sample and iterate this process over all samples in the given data set). Three experiments were conducted where the gender strength variable and the ill-defined samples

[3]Matlab implementation of LDA and SVM is used.

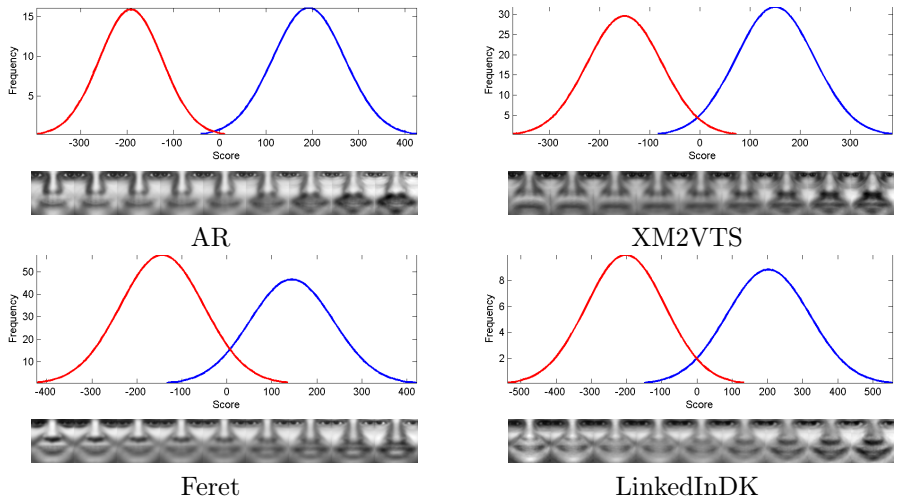


Figure B.3: The estimated gaussian distribution of the data sets projected to the GSR projection line, with synthesized faces corresponding to points on the regression line.

were estimated by (case 1) only the time-limited test, (case 2) only the non-time-limited test and (case 3) by both the time-limited and non-time-limited test combined. The results are presented in Table B.3. As the dimension of the variables is much higher than the samples for all data sets, a dimensionality reduction by Principal Component Analysis (PCA) [95] was preformed prior to building the models. The PCA was set to retain 95% of the data set variance.

B.5.3.2 Generalization evaluation

To see how the algorithms perform on images not included in four data sets, an experiment was conducted, where all four data sets were collapsed into one training set of 1886 images and these were tested on a data set consisting of 40,692 images from public Danish LinkedIN profiles (LinkedInDKFULL). In this experiment we used a radial basis function kernel for SVM. Finding the optimal parameters for a non-linear SVM is still an open problem, therefore we employ a gridsearch strategy to determine the C and γ, in the grid $[10^{-2}10^{5}]$ and $[10^{-5}10^{1}]$, respectively. It was found that C $= 10^4$ and $\gamma = 10$ was the optimal setup. The results are presented in Table B.4, where it can be seen that removing ill-defined training samples had a huge impact.

Full training set

Case 1

Data set	LDA	SVM	GSR
AR	0.0604	0.0604	**0.0570**
XM2VTS	0.1304	0.1304	**0.1250**
Feret	**0.1364**	**0.1364**	0.1495
LinkedInDK	**0.1950**	0.2000	0.2000

Case 2

Data set	LDA	SVM	GSR
AR	**0.0604**	**0.0604**	0.0638
XM2VTS	0.1304	0.1304	**0.1232**
Feret	0.1364	0.1364	**0.1304**
LinkedInDK	**0.1950**	0.2000	0.2050

Case 3

Data set	LDA	SVM	GSR
AR	0.0604	0.0604	**0.0570**
XM2VTS	0.1304	0.1304	**0.1268**
Feret	**0.1364**	**0.1364**	0.1376
LinkedInDK	**0.1950**	0.2000	0.2100

Refined training set

Case 1

Data set	LDA	SVM	GSR
AR	**0.0537**	**0.0537**	0.0671
XM2VTS	0.1286	0.1286	**0.1250**
Feret	**0.1280**	0.1292	0.1388
LinkedInDK	**0.1800**	**0.1800**	0.2000

Case 2

Data set	LDA	SVM	GSR
AR	**0.0604**	**0.0604**	0.0839
XM2VTS	**0.1196**	**0.1196**	0.1304
Feret	**0.1244**	0.1256	0.1292
LinkedInDK	**0.1900**	**0.1900**	0.1950

Case 3

Data set	LDA	SVM	GSR
AR	**0.0570**	**0.0570**	0.0638
XM2VTS	**0.1159**	**0.1159**	0.1232
Feret	0.1232	**0.1184**	0.1352
LinkedInDK	**0.1700**	**0.1700**	0.2100

Table B.3: FCR obtained by a leave-one-out cross-validation scheme for case 1 to 3, including (full training set) or excluding (refined training set) ill-defined training examples. Bold signifies lowest FCR for a data set in the corresponding test and green signifies lowest FCR for a data set between tests in (full training set) and (refined training set).

(a)

Data set	LDA	SVM	GSR
LinkedInDKFULL	0.3762	**0.3227**	0.3521

(b)

Data set	LDA	SVM	GSR
LinkedInDKFULL	0.2520	**0.2172**	0.2373

Table B.4: FCR obtained by training on all four data sets and testing on 40,692 public Danish LinkedIN profile pictures, including (a) or excluding (b) ill-defined training examples. Bold signifies lowest FCR for a data set in the corresponding test and green signifies lowest FCR for a data set between tests in (a) and (b).

B.6 Discussion and Conclusion

In this work we have presented the gender strength variable, a new way of determining class membership that supersedes the conventional binary classes normally used in gender recognition. We have assessed the benefit of the gender strength variable by identifying and removing ill-defined training samples to improve existing methods such as LDA and SVM. Employing the refined training set on LDA and SVM (both in a linear and non-linear version) gave a significant improvement. We could see a performance increase when pooling answers from both cognitive tests (with and without time constraints) probably due to the fact that we obtained a more robust estimation of ill-defined training samples. This indicates that not only the volume of the training data is important but also the quality of the training data. However, how to optimally determine ill-defined training samples is still an open question.

Interestingly we also observed an improvement with SVM. As SVM manly focuses on samples close to the decision boundary one could expect that removing ill-defined samples would lead to a decrease in performance. However, here we should remember that the two populations (male and female) cannot be perfectly separated, and ill-defined samples my not lie adjacent to the decision boundary, which results in the observed performance increase.

A new method was also devised - the Gender Strength Regression. Results indicate that without removing ill-defined training samples GSR performs similarly to or better than the LDA and SVM (in a linear version). However, when using the refined training set, LDA and SVM outperforms GSR. Also the complexity of the data seems to be best modeled by non-linear methods, however these

methods need more tuning of parameters.

Human performance was reported, to the authors' knowledge for the first time, on the four data sets. When comparing the performance of human vs. automatic machine learning algorithms, it can be seen that the machine learning algorithms outperform humans on the AR data set. Machine learning algorithms perform only 1-3% worse than humans on the Feret and XM2VTS data sets. But when going from the relatively constrained data sets (similar frontal poses) to the more unconstrained data set of Danish LinkedIN profile pictures, humans outperform the machine learning algorithms. It should be kept in mind that all tests were performed on the central face in gray scale images.

Finally, the results obtained for the four data sets were validated by reproducing the trend that a performance boost was achieved when removing ill-defined training samples, by collapsing the four training sets and testing them on a data set of 40,692 public Danish LinkedIN profile pictures.

It should be noted that the gender strength variable seems to be well estimated by a linear encoding of the gender classification response times, however it is not known if this is the optimal encoding.

Appendix C

3D Gender recognition using cognitive modeling

Jens Fagertun, Tobias Andersen, Thomas F. Hansen and Rasmus R. Paulsen

Abstract

We use 3D scans of human faces and cognitive modeling to estimate the *"gender strength"*. The *"gender strength"* is a continuous class variable of the gender, superseding the traditional binary class labeling. To visualize some of the visual trends humans use when performing gender classification, we use linear regression. In addition, we use the gender strength to construct a smaller but refined training set, by identifying and removing ill-defined training examples. We use this refined training set to improve the performance of known classification algorithms. Results are presented using a 5-fold cross-validation scheme and also reproduced using an unseen data set Keywords: *3D gender recognition, Linear Discriminant Analysis, Support Vector Machines, Cognitive Modeling, Linear Regression.*

C.1 Introduction

Determining gender is a natural, subconscious task performed every day in human interaction. By continuously training this task, the human mind is shaped

to become a gender recognition expert. This gives rise to some interesting questions: "Can we understand how humans perform the task of gender recognition?" and "Can we use this information to enhance machine learning models?" These questions have previously been addressed using 2D image data [52]. Here we applied the same question to the 3D domain, taking both shape and texture into consideration and using 3D facial scans.

The question of gender classification is not only interesting in itself; it can also be used to improve performance of classification algorithms working on biometrics that contain identity (face, iris etc.) as described in [152].

When confined to non-intrusive data (such as visual 2D/3D or audio data), the two populations (male and female) cannot be perfectly separated. There are border-line cases where determining gender can be extremely difficult for humans; cases where sports athletes need to undergo a DNA test to determine gender, come to mind.

The problem of automatic gender classification has been studied for years. Many algorithms have been proposed that employ different methods, including Linear Discriminant Analysis (LDA) [11], Haar-like wavelets [106], Local Binary Patterns (LBP) [106] and Support Vector Machines (SVM) [114]. All these methods work on a binary understanding of class association, which does not necessarily enable the method to identify and exclude ill-defined samples, and this could lead to a performance decrease.

Obviously, humans perform gender recognition based on a variety of data, such as face, hairstyle, clothes, height, voice, and movement patterns. In this work we look exclusively at faces. To address the previously posed questions, this work estimates a continuous "gender strength variable" for individual facial scans based on human performance. This gender strength variable and linear regression is then used to visualize some of the visual trends humans use when performing gender classification. In addition, we use this gender strength variable to construct a smaller but refined training set, by identifying and removing ill-defined training examples. We then assess if this refined training set improves performance of known classification algorithms.

The structure of the paper is as follows. Section C.2 introduces the 3D data used in this work. Section C.3 describes the cognitive test that estimates the gender strength variable. Section C.4 presents the Gender Strength Regression method. Section C.5 presents the experiments and the obtained results. Finally, Section C.6 presents a discussion and conclusion on this work.

C.2 Data sets

In this work we use 720 3D facial scans of unrelated subjects recruited among volunteers in the Danish Blood Donor Study [122]. The facial scans were captured using a Canfield Vectra M3 Imaging System and have been cropped to only show the face, excluding neck, ears and hair. The 3D facial scans contain shape information (x-, y-, z-point positions) and texture information (red, green, blue intensities in every 3D point). The cropped 3D facial scans have been remeshed so that all 3D points have a point-to-point correspondence between the individual scans using a method similar to [16] and then the scans have been aligned by a Procrustes analysis [69].

The data is divided into two sets; DATA I (n=396) for use in the cognitive test and DATA II (n=324) for an unseen set to test the results of DATA I. DATA I includes the same number of males and females. This should motivate the participants (in the cognitive test) to employ a response criterion that is not biased towards one of the two response categories. A sample of the 3D data is shown in Figure C.1.

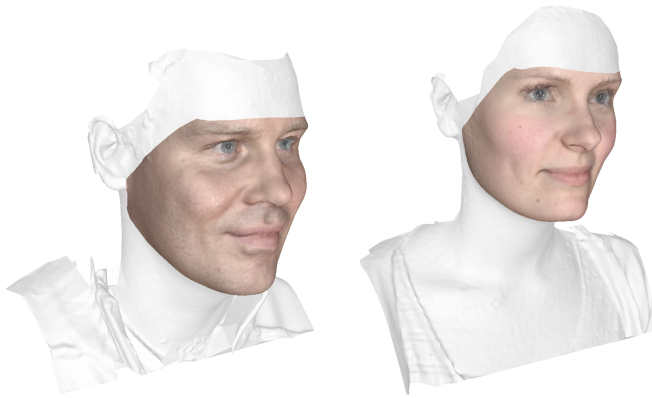


Figure C.1: Example of the 3D data used. Light gray parts are discarded.

C.3 Cognitive test

The object of the cognitive test is to estimate the gender strength of each sample (facial scan). The gender strength variable is defined as a variable within the interval [0 1], where close to zero signifies belonging poorly to the sample's class,

whereas 1 signifies belonging strongly to the sample's class. This is similar to what is done in fuzzy sets [167]. To encode the gender in the variable, females will be denoted in the negative interval [-1 0] and males in the positive interval [0 1].

In order to estimate the gender strength variable, a cognitive user test was devised and conducted by six test participants. The gender strength variable is estimated as a latent variable encoded by the time it takes participants to perform the gender classification task.

The cognitive test is conducted with the psychtoolbox for matlab [19], to ensure precisely recorded answer times. In the test, every facial scan is shown as a 2D image showing the frontal face and the two profiles, see an example in Figure C.2. The time it takes the test participant to make the classification is recorded. A graphical illustration of the tests is shown in Figure C.2. Six participants conducted the test, where the order of stimuli was randomized to avoid order effects.

The gender strength variable is estimated by standardizing the test participants´ answer times, so that all answer times from one test participant have mean zero and unit variance. Then, a pooling of the standardized answer times from all test persons is performed. Finally, the pooled answer times below the mean are encoded to one, the outliers (above three standard deviations) are encoded to zero and the remaining are linearly encoded between the values zero and one.

C.4 Gender Strength Regression

In classification problems, training data is usually only labeled with the class to which it belongs. As a result, classifiers, such as LDA and SVM, have only a binary understanding of class membership, which they use in the objective function to construct the classification method. LDA seeks to minimize the within-class variance, while maximizing the between-class variance, whereas SVM tries to maximize the margin (distance) between classes. Due to this binary understanding of class, these methods work on the complete data set. The performance of methods such as LDA and SVM depends solely on the given training data, where more training data does not necessarily yield a better performance.

In this paper we suggest enhancing the performance of classifiers such as LDA and SVM by removing ill-defined training samples. In this work, ill-defined samples are defined as cases where more than 50% of the test participants in the human cognitive tests fail to perform a correct classification.

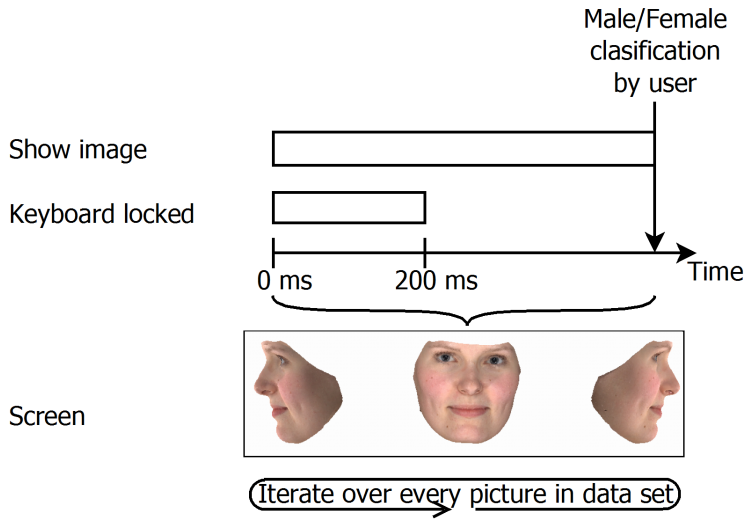


Figure C.2: Timeline for the cognitive test. There is a 550 ms delay from a user classification to when the next picture is shown.

Furthermore, as in [52], we use a classifier that seeks to estimate the gender strength variable by regression, Gender Strength Regression (GSR). A linear least squares regression has the objective function

$$\min_{\beta} \|\beta \mathbf{X} - \mathbf{y}\|^2, \tag{C.1}$$

with the closed-form solution

$$\beta = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}, \tag{C.2}$$

where $\mathbf{X}$ is a matrix with samples and variables in the rows and columns, respectively; $\mathbf{y}$ is a vector with the gender strength scores, also called the response variable; and β is the coefficient for the regression line. In an optimal system ((C.1) is small) the decision threshold would be zero in accordance with how we encoded the response variable in the regression, see Section C.3. However, this will not always be the case. A more accurate decision threshold can be calculated by

$$\frac{\sum_{i=1}^{k} \beta \mathbf{x}_i}{2k} + \frac{\sum_{j=1}^{l} \beta \mathbf{x}_j}{2l}, \tag{C.3}$$

where i and j belong to samples of class male and female, respectively; and k and l are the sizes of the two classes. This assumes that the two classes have the same distributions.

C.5 Experiments

C.5.1 Cognitive test

In the cognitive test, the False Classification Rate (FCR) is reported in Table C.1. Interestingly, there seems to be a clear difference between male or female being wrongly classified.

Human False Classification Rates	
Male wrongly classified as a female	$0.84\% \pm 0.44\%$
Female wrongly classified as a male	$6.65\% \pm 3.01\%$
Total	$\mathbf{7.49\% \pm 2.85\%}$

Table C.1: FCR for the cognitive test with standard deviations.

C.5.2 Projection lines

To obtain an understanding of how GSR classifies gender, all samples in the data sets are projected onto the respective GSR line. As opposed to the 2D test [52] that visualizes the features understood in the first 200ms of exposure, we do not have a time limit so that the visualized features are not necessarily the "first" understood by the observer.

Synthesized faces between the female and male extremum points on the GSR regression line are shown in Figure C.3 for both shape and texture.

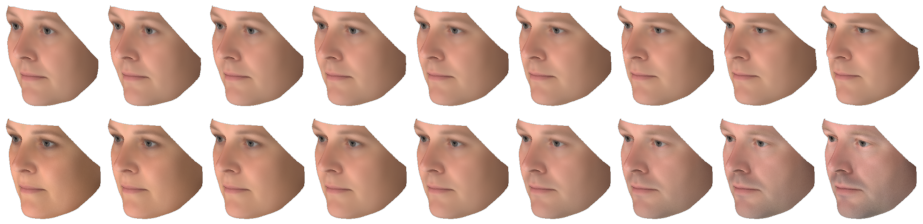


Figure C.3: Synthesized faces between the female and male extremum points on the GSR regression line. Shape and texture is presented in the first and second row, respectively. With a female representation to the left, a male representation to the right and the mean shape/texture in the center.

C.5.3 Recognition results

C.5.3.1 Cross-validation evaluation

Gender classification was performed for DATA I using GSR and compared to LDA and SVM with a linear kernel in a 5-fold cross-validation scheme. (Train on four folds of the data. Then test on the "unseen" fold and iterate this process over all folds.) The results are presented in Table C.2, in which we have separate tests for shape and texture to see how much discriminating power they individually contain.

As the dimension of the variables is much higher than the samples for all data sets, a dimensionality reduction by Principal Component Analysis (PCA) [95] was preformed prior to building the models. The PCA was set to retain 95% of the data set variance.

Training set size		LDA	SVM	GSR
Shape	Full set	4.80%	4.80%	5.30%
	Refined set	4.29%	4.55%	5.30%
Texture	Full set	1.77%	1.77%	1.52%
	Refined set	**1.01%**	**1.01%**	2.02%

Table C.2: FCR obtained by a 5-fold cross-validation scheme including (full training set) or excluding (refined training set) ill-defined training examples. Bold signifies lowest FCR.

C.5.3.2 Unseen evaluation

A second test was conducted where DATA I was used for training and DATA II was used for the test. In this experiment we used a radial basis function kernel for SVM. Finding the optimal parameters for a non-linear SVM is still an open problem, therefore we employ a grid-search strategy to determine the C and γ , in the grid $[10^{-2}\ 10^{5}]$ and $[10^{-5}\ 10^{1}]$, respectively. It was found that $C = 10^{-2}$ and $\gamma = 10^{1}$ for shape and $C = 10^{3}$ and $\gamma = 10^{2}$ for texture was the optimal setup. The results are presented in Table C.3.

Training set size		LDA	SVM	GSR
Shape	Full set	6.17%	5.25%	6.48%
	Refined set	4.94%	4.63%	4.94%
Texture	Full set	2.16%	2.16%	1.85%
	Refined set	1.85%	**1.54%**	2.16%

Table C.3: FCR obtained by training on DATA I (full training set or refined training set) and testing on DATA II. Bold signifies lowest FCR.

C.6 Discussion and Conclusion

In this work we have assessed the benefit of the gender strength variable by identifying and removing ill-defined training samples to improve existing methods such as LDA and SVM on 3D facial scans. Employing the refined training set on LDA and SVM (both in a linear and non-linear version) gave an improvement that is most prominent in the texture experiments. Like in [52], this indicates that not only the volume of the training data is important but also the quality of the training data. However, how to optimally determine ill-defined training samples and, more interestingly, how to identify them without an intensive user interaction, is still an open question.

Interestingly we also observed an improvement with SVM. As SVM mainly focuses on samples close to the decision boundary, one could expect that removing ill-defined samples would lead to a decrease in performance. However, here we should remember that the two populations (male and female) cannot be perfectly separated, and ill-defined samples may not lie adjacent to the decision boundary, which might result in the observed performance increase.

When looking at the shape and texture experiments it is clear that texture holds much more discriminating power than shape.

Using 3D facial scans, we obtain a difference between human and machine performance, larger than the ones observed for 2D [52]. This is probably due to a variety of factors. First we here have a point-to-point (and texture intensity to texture intensity) correspondence so that a considerable amount of noise from variations in shape is filtered out of the texture domain. The facial scans used here have a considerably larger "resolution" than the 2D images in [52], and finally we have more information in a 3D surface than in 2D plane so we would expect a performance increase.

The result obtained indicates that when working with high resolution 3D facial scans, standard machine learning algorithms such as LDA and SVM out-perform humans on data acquired in a constrained environment. If this also remains valid in more unconstrained environments remains to be seen.

C.7 Acknowledgement

We thank the participants in the Danish Blood Donor Study, in particular Henrik Ullum. This project was sponsored by a grant to Thomas Werge from The Danish National Advanced Technology foundation (Jnr. 001-2009-2) and Lundbeck Foundation (R34 A3243) and by grants to Thomas Hansen from the Danish Main Capital Research Foundation of Health. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Appendix D

Variability of 3D facial landmarks

Jens Fagertun, Stine Harder, Anders Rosengren, Christian K. Møller, Thomas Werge, Rasmus R. Paulsen and Thomas F. Hansen

Abstract

A full facial 3D annotation procedure is presented for a sparse set of manually annotated landmarks. Manual annotation of landmarks is a known source of variance, which exist in all fields of medical imaging, influencing the accuracy and interpretability of the results. However, the variability of human facial landmarks is only sparsely addressed in the current literature as opposed e.g. to the research field of orthodontics and cephalometric.

In our approach, we analyze an extended set of manually annotated landmarks in regards to intra- and inter-operator variability. We then select a sparse set of landmarks in order to construct dense correspondence of the 3D scans with a minimum of point variance. Using the sparse set of landmarks (n=14), we minimize the point mean variance. Further, we provide a 3D full facial map as a reference for annotation variation of other putative landmarks using this sparse landmark set. Keywords: *3D Facial Landmarks, Inter-Operator Annotation Variance, Dense Point Correspondence, Point Distribution Model, ANOVA.*

D.1 Introduction

The research field of facial morphology has advanced rapidly over the last ten years, with the introduction of better, faster and cheaper systems for facial 3D scanning. The systems have enabled more accurate and objective methods of capturing facial morphological differences. Analysis of facial morphology is based on facial distances i.e. the distance between facial landmarks [29, 102, 119] or on statistical models [29, 74]. One widely used statistical method, uses Principal Component Analysis (PCA) to assess the population's variance and is referred to as a Point Distribution Model (PDM) [86]. Both methods rely on manually annotated landmarks that are used directly or as a basis for constructing a dense point correspondence [121, 86, 74, 29]. This means that both direct distances and statistically based methods are prone to human operator annotation errors. While there exist a great number of surface-based automatic registration methods for point correspondence, manual annotation, at least on a sparse set of landmarks, is widely used when facial analysis is used in clinical applications. Understanding the variance (noise) introduced by manually annotated landmarks is important for the design and interpretation of studies in this field.

The reliability of facial landmark annotation has not been as thoroughly studied as landmark annotations in other fields. For example, Buschang *et al.* [24] assessed the inter-operator annotation variability of anatomical landmarks on the skull for use in orthodontics and cephalometric analysis, using an ANOVA analysis. Similar, newer studies also addressed the reliability of cranial-anatomical landmarks [159, 140, 8]. In [97] the inter-operator annotation variance was included in the PCA when analyzing cranial growth. Here the landmark variance was addressed using a weighting scheme giving most weight to annotation landmarks with low variance.

In this work, we exclusively work with human facial features. We address the reliability when annotating facial features with respect to inter/intra operators and samples (portraits). To the best of our knowledge, no other studies have addressed this in the field of facial morphology. On the basis of the reliability of facial landmarks, we suggest a sub-set of landmarks that yields a superior dense-point correspondence compared to the original landmarks.

D.2 Methods

D.2.1 Sample and image data

The data used in this work consists of 36 facial scans of unrelated healthy subjects recruited among volunteers in the Danish Blood Donor Study (DBDS) [122]. The 36 subjects were randomly chosen from our database consisting of facial scans from 641 subjects, and they were evenly divided between men and women. The facial scans were captured using a Canfield Vectra M3 Imaging System, at the DBDS facility at Glostrup University Hospital. Each 3D facial scan contains about 70,000 to 100,000 3D points and has shape information (x-, y-, z-point positions) and texture information (red, green, blue intensities) for every 3D point.

D.2.2 Description of annotations points

The annotation framework initially developed in Fagertun [50] consists of 73 landmarks. Here, 24 anatomical landmarks define distinct facial features, and 49 pseudo-landmarks define the curves and width of the jaw, lips, eyebrows etc. A description is presented in Figure D.1.

D.2.3 Annotation procedure

All scans followed a three-step annotating scheme:

1. Automated annotation of landmarks, detailed description in D.2.4:
    - A fully automatic Active Appearance Model (AAM) [34] in 2D.
    - An Active Shape Model (ASM) [37] in 3D.
2. Correction by human operator of the pre-annotated landmarks, detailed description in D.2.5.
3. Post processing, detailed description in D.2.6
    - Creation of dense point correspondence meshes.

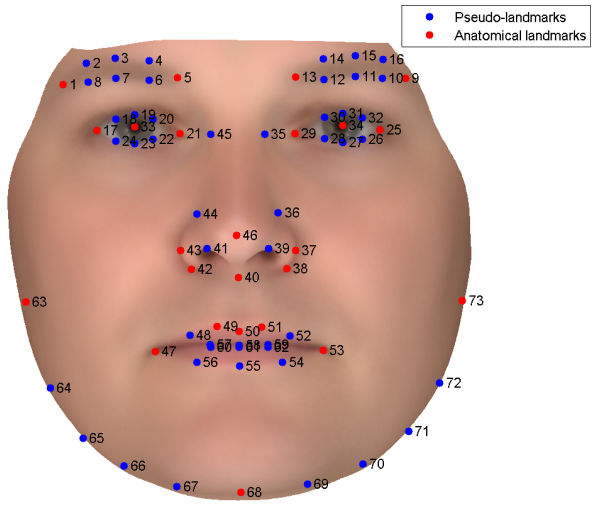


Figure D.1: Facial landmarks divided into anatomical and pseudo-anatomical landmarks.
Anatomical landmarks: 1) Right eyebrow lateral point, 5) Right eyebrow medial point, 9) Left eyebrow lateral point, 13) Left eyebrow medial point, 17) Right eye lateral canthus, 21) Right eye medial canthus, 25) Left eye lateral canthus, 29) Left eye medial canthus, 33) Right pupil center, 34) Left pupil center, 37) The outer left alar-facial groove, 38) The inner left alar-labial groove, 40) The columellars connection to the upper lip, 42) The inner right alar-labial groove, 43) The outer right alar-facial groove, 46) Tip of the nose, 47) Right oral commissure, 49) The right philtrum column connection to the the vermillion border, 50) Midpoint on the cupid's bow, 51) The left philtrum column connection to the the vermillion border, 53) Left oral commissure, 63) Right ear attachment, 68) Lowest point of the central jaw, 73) Left ear attachment.
Pseudo-landmarks groups: 2-4 & 6-8) Right eyebrow, 10-12 & 14-16) Left eyebrow, 18-20 & 22-24) Right eye, 26-28 & 30-32) Left eye, 35 & 36 & 39 & 41 & 44 & 45) Nose, 48 & 52 & 54-62) Mouth, 64-67 & 69-72) Jaw.

D.2.4 Data pre-processing by automatic annotations

A 2D image was created by orthographic projection of the 3D scan. The face and eyes are automatically detected by a Viola-Jones Rapid Object Detection [158, 157], and serve as a starting point for an AAM search. When the AAM converges, the 73 2D annotation points (Figure D.1) can be extracted. These annotation points are then transformed from the 2D image to the 3D scan. The 2D to 3D transformation is likely to fail in high curvature areas like the jaw (points from 2D are wrongly projected onto the neck). To compensate for this limitation, an ASM search, initialized by an Iterative Closest Point [169] search, is performed to locate the jaw in 3D. The annotation points are then corrected by an operator as described in the next section. In summary, the low curvature points are found by a 2D AAM and transformed to 3D, while high curvature points are found by a 3D ASM.

The 2D AAM and 3D ASM where constructed based on 605 individuals recorded by a Nikon D90 in 2D and a Canfield Vectra M3 Imaging System in 3D. Both the 2D and 3D data was annotated to create correspondence between individuals, in the same fashion as described in the following section.

D.2.5 Manual annotation tools and standard

The object of the manual annotation was to reach a consistent and stable standard for annotation. The annotation scheme was explained and discussed during a three-hour workshop, where a common frame of reference was reached. The manual annotation is a two-step process. Firstly, the annotation is performed in a fixed frontal view by a custom-made annotation tool. Secondly, points in high curvature areas are adjusted in fixed frontal, profile, top/down views. The high curvature points are the jaw and nose points (35-45 and 63-73). Pilot analysis showed that using a 3D annotation program, with possibility to rotate and zoom the scan, was very time consuming and did not improve error rate when compared with the chosen procedure, data not shown.

D.2.6 Dense point correspondence

To analyze facial shape variation at positions not annotated by landmarks, a dense point correspondence is created. A variety of methods exist for establishing dense correspondence. In this work we employ a method that has previously produced excellent results when a sparse set of landmarks exist [121].

This method is based on propagating a well-formed template mesh to all shapes in the training set. For each shape the template mesh is initially deformed using a volumetric thin-plate spline warp [18] and using the sparse set of corresponding landmarks. In the next step the mesh vertices of the deformed template mesh are propagated to the target shape. This approach is very similar to the method used to create the dense surface models described by Hutton *et al.* [86, 74, 29]. While propagating each vertex to the Euclidian closest point on the target surface works for simple anatomies, it fails in regions with moderate curvature. A proven solution is to regularize the correspondence field and add curvature information in the propagation step. In [121, 82] this regularization is cast into a Markov Random field (MRF) framework [100], where a prior and an observation term are defined. The prior model imposes a Gaussian prior on the deformation field that favors smooth deformation fields. The curvature of the deformed template mesh and the target shape is used in the observation term to guide the correspondence to areas with similar curvature. The mean curvature is estimated as the radius of a locally fitted sphere [120]. Finally, the regularization is bounded so the projected points are on the surface of the target shape. The optimal correspondence field is found using stochastic optimization. The involved weighting between the prior and observation terms is found as the weight that creates the most compact shape model as described in [82]. The result is a regularized dense correspondence between the template and all the shapes in the training set. In our experiments, the dense correspondence consists of 39,653 points and the associated mesh connectivity from the template mesh.

D.3 Results

D.3.1 Landmark variability

Six operators(one female) annotated 36 scans (of 18/18 females/males, aged 18 to 65) twice, one week apart. The mean error and standard deviation of the

combined variance of each annotation point are shown in Figure D.2. It is clear that the variability is dependent on the annotation point in question. The center of the pupil was associated with minimal variance (SD = 0.09mm), followed by landmarks of the eye (SD = 0.30-0.95mm). The most error-prone annotation points are the landmarks of the jaw (SD = 1.55-3.34mm), although the lateral points of the eye-brows are also error prone (SD = 2.24-2.37mm). The variance of each annotation point is illustrated in Figure D.3.

D.3.1.1 Intra/inter operator variability

We used a mixed-model ANOVA analysis, using the Minimum Variance Quadratic Unbiased Estimation (MIVQUE) method to estimate the effects of the components: operator, session day, and the scan number (portrait), for each of the 73 annotated points:

$$Y_{ijk} = \mu + O_i + D_j + I_k + \epsilon_{ijk} \tag{D.1}$$

where Y_{ijk} is the data sample, μ is the global average, O_i, D_j and I_k is the main effect term for inter-operator, intra-operator (session day) and portrait, respectively. ϵ_{ijk} is the error term for unexplained variance. A three-way ANOVA using interaction terms was rejected as the model did not contribute with further explanation of the variance, data not shown.

Generally the session day, thus intra-operator, contributed relative little to the variability (SD <= 0.25mm), see Figure D.4. The most reliable annotation landmark was the center of the pupil, as this was only marginally influenced by the inter/intra-operator and portrait, and was not associated with a large error term. There were no significant differences between landmarks and pseudo-landmarks. Still, the variance was more prominent in the points describing the jaw and nose and to some extent the medial canthus of the eye, Figure D.4.

D.3.2 Dense point correspondence optimization

As can be seen, the 73 annotated points were associated with different variability. We test four different sub selections of these annotation points in a effort to minimize the variance of the resulting dense point correspondence. Two landmark selections simply exclude annotation points with the high variance in mean error (>1mm) and operator error (>0.5mm), respectively. However, these

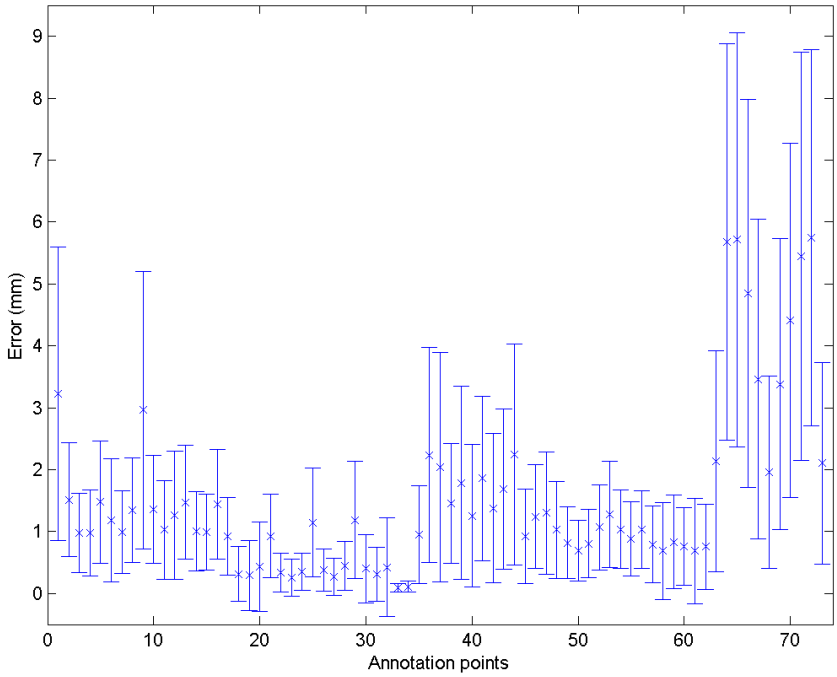


Figure D.2: Mean annotation error.

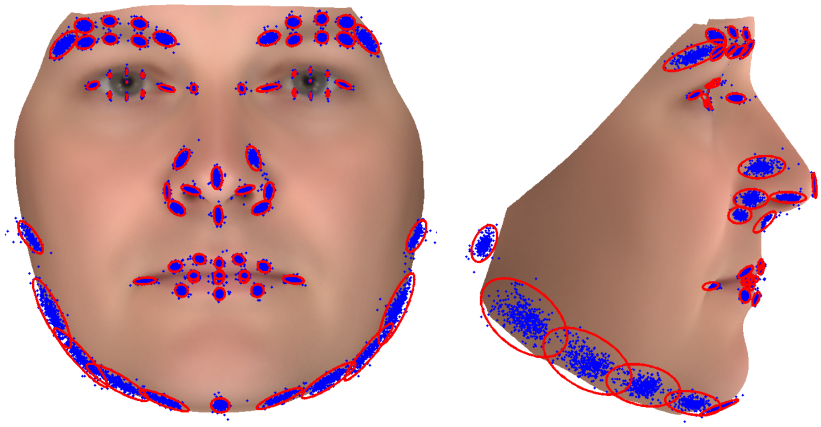


Figure D.3: Frontal and profile annotation variance plot (Red ellipse shows 3 SD).

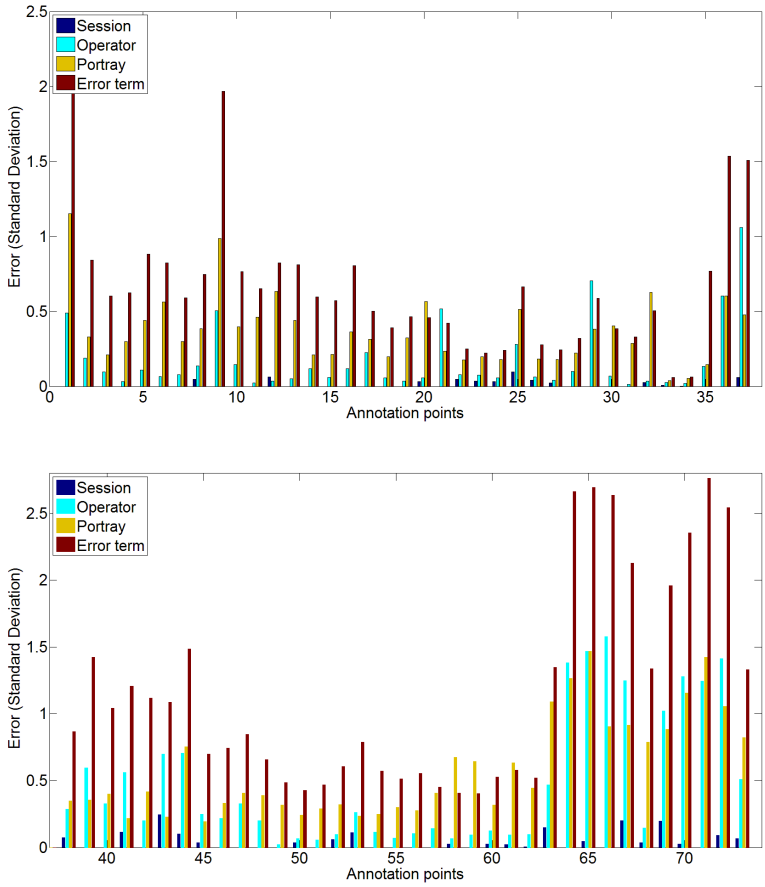
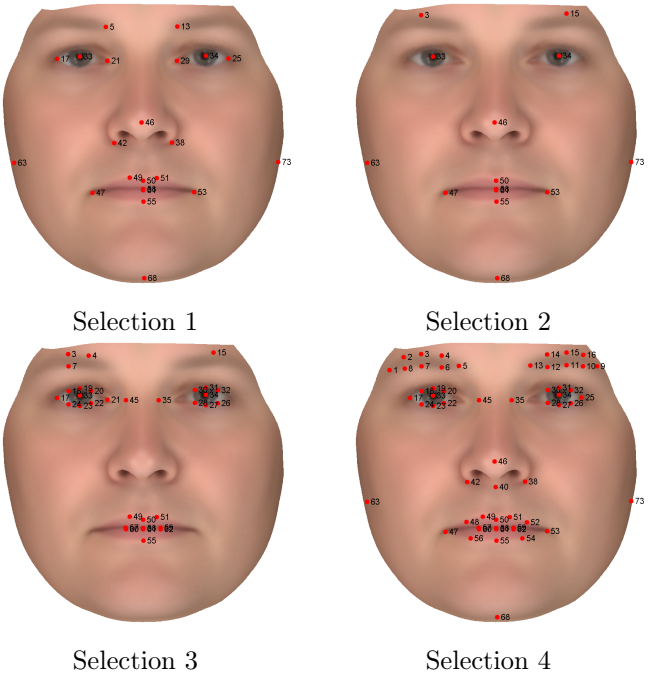


Figure D.4: The variance of the operator, session day and portrait, respectively for each of the 73 annotated points.

landmark selections do not take into account whether landmarks from the whole facial area are selected. Thus, additionally two landmark selections aiming at selecting landmarks from the main facial features and with low mean error and variance are also tested, see Figure D.5.



Selection 1 Selection 2

Selection 3 Selection 4

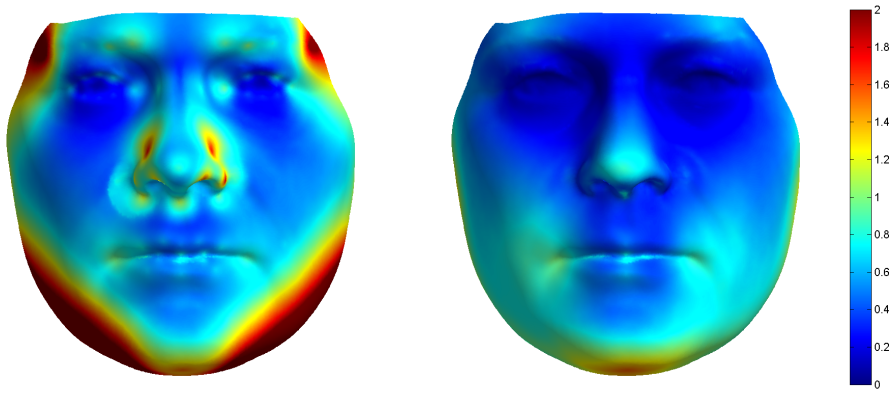
Figure D.5: Landmark selections 1,2,3, and 4. Selections 1 and 2: Selected for "having low mean error and variation" with focus on capturing all facial features. Selection 3: Mean error below 1mm. Selection 4: Operator standard deviation below 0.5mm.

The quality of the derived dense point correspondence was evaluated using the Minimum Description Length (MDL). The Minimum Description Length (MDL) principle states, in an Occam's Razor way, that the most compact model will be the best to express the data, as it has the best possible correspondence between points. If the points have good correspondence, the cumulative variation in the PDM will increase quickly over the first principal components. In the case of bad correspondence, the variation will increase slowly, as the model has to model both the population variance and the inconsistency of correspondence points. This approach has previously been used by Davies and colleagues [40, 41], were the MDL serves as an objective function for automatically generating landmark correspondence on 2D shape outlines or 3D surfaces.

Landmark selections	Full	1	2	3	4
Mean variance	1.92	0.62	**0.54**	1.03	0.71

Table D.1: Dense point mean variance.

We use the MDL measure to evaluate the quality of the point correspondence of the different landmark selections. We measure the MDL of the dense mean point variances for all the annotation points and the four suggested landmark selections and the results are shown in Table D.1. The lowest variance is seen for landmark selection 2, having a mean variance = 0.54. Figure D.6 visualizes variation of the PDM from landmark selection 2 compared to the original PDM (based on all landmarks). Table D.2 shows that points with a relatively small variation in the user annotations cannot be estimated better automatically. However, points with big variation in the user annotations can indeed be better estimated. This yields the conclusion that a reduced set of the original 73 landmarks is preferred. It is also noted that landmark selection 2, that consists of 14 landmarks, results in a more compact model and a better estimation of 16 of the 73 landmarks compared to the manual operator annotations, data not shown.



All points Landmark selections 2

Figure D.6: Visualization of variance in mm (of one standard deviation).

Annotation point number	User	All 73 points	Landmark selection 2
33 (Right pupil center)	**0.093**	0.205	0.170
34 (Left pupil center)	**0.112**	0.406	0.391
50 (Midpoint on the cupid's bow)	**0.694**	0.710	0.708
47 (Right oral commissure)	1.299	1.300	**1.296**
53 (Left oral commissure)	1.281	1.261	**1.260**
66 (Point on the right jaw)	4.851	4.789	**3.754**
70 (Point on the left jaw)	4.411	4.383	**3.334**

Table D.2: Mean error for selected points and landmark selections.

D.4 Discussion and Conclusion

To the best of our knowledge, this is first study to address the variation of human-annotated 3D facial landmarks. Understanding the variation of manual annotations is important as much work in registration, recognition, and machine learning is influenced by manual annotation errors. However, the current literature is sparse when it comes to 3D facial morphology and variation. We expect that an increase in accuracy, user friendliness (less operator demanding) and availability of 3D imaging scanners will probe the use of shape models in clinical diagnostics, as seen e.g. for orthopedic surgery [148]. However, to assess the putative clinical impact of such tools, it is important to understand the variability embedded in manual annotation. We focus on facial morphology, and suggest a procedure to retrieve a dense correspondence mesh of the face with low variance and minimum of human operator annotation points.

We first address the variability of 73 facial 3D landmarks, and that the variability is highly correlated with the annotation point, per se. As expected landmarks that are easier to define in consensus have the lowest variability concerning both inter- and intra-operator variability. As expected, it is the landmarks of the pupils that are associated with the minimum variation. More leniently defined landmarks such as the points defining the jaw line are associated with the highest variation. We repeated the annotation twice, and although we believe that further repeating of the annotation will not change the variation notably, we cannot exclude that it will change.

One obvious application of the annotated points is to identify minor facial abnormalities, e.g. to assist diagnosing syndromes. Such abnormalities can be identified by using absolute measures or the ratio between manually annotated landmarks, or by using a dense correspondence mesh. It is obvious to use a dense correspondence mesh for identification of minor abnormalities, as this fa-

cilitates the use of landmarks/points not manually annotated and thus a larger data set. In a clinical setting, different operators will be used, and although such operators will be ideally trained, the variability will influence the results. We suggest an approach to limit the inter (and intra) operator number of annotation points to enable capture of facial features and a minimum of variability. The model uses 14 landmarks to create a dense correspondence mesh with a point mean variance of 0.54. Further, the model shows less variability in 16 of the manually annotated points not included in the model. Using fewer annotation points will decrease the operator time, thus making it more feasible to use.

There is one limitation with regard to generalizability of the study. We used subjects that are Caucasian with Scandinavian background, thus we cannot exclude that the variability of the annotation landmarks is different from other ethnicities, e.g. the texture of blonde eyebrows on light skin may be difficult to separate, whereas dark eyebrows may not.

Appendix E

Correlation of complex polygenic variations to specific holistic facial characteristics

Jens Fagertun, Tune Pers, Karin Wolffhechel, Henrik B. Nielsen, Daniel Gudbjartsson, Hreinn Stefansson, Kári Stefansson, Rasmus R. Paulsen and Hanne Jarmer

Abstract

The effect of the human genome in the context of facial appearance is an emerging field receiving increasing attention. Specific single nucleotide polymorphisms have been connected to simple facial distances. In contrast to such previous studies, we attempt a holistic approach correlating complex polygenic variations to holistic facial characteristics across a population of 1,338 individuals. The focus is to find a one-way mapping between a genome-wide SNP model and a facial statistical shape model. We show that statistical SNP models can indeed recover some of the heredity present in facial characteristics, and for the best performing facial components, we are able to classify to simple facial archetypes better than would be expected by chance.
Keywords: *Facial analysis, Facial characteristics, Genome-wide association studies, Eigenshapes, Eigengenes, Single nucleotide polymorphisms.*

Introduction

Apart from being nearly genetically identical, monozygotic twins are often so similar in appearance that they can be difficult to tell apart. Their similarities are believed to be due to combinations of genetic, epigenetic and environmental factors [7]. Genotypic profiling based on single nucleotide polymorphisms (SNPs) has been successfully used to predict facial appearance, including the amount of freckling, the presence of moles, hair texture and skin color [149, 94, 49, 110, 147, 48, 92]. In forensic science genotyping has been used to predict an individual's hair color, eye color, sex and ancestry with high accuracy [93, 4]. Especially the prediction of facial appearance based on genotype data holds great promises in forensic science, because this would allow re-construction of an individual's facial appearance, which could aid in identifying crime suspects.

However, recent genome-wide association studies (GWAS) have shown that most traits are inherently polygenic [81, 96, 107]. One example is human height, a trait that is believed to be shaped by thousands of SNPs [166]. Although facial traits are arguably less complex than the human height trait, it is likely that they are the product of hundreds of genetic variants. Recently, three SNPs affecting nose width, nasion position and face width have been detected [17, 119]. Liu *et al.* attempted to link genetic variants from close to ten thousand individuals of European descent to facial shape phenotypes extracted from three dimensional (3D) magnetic resonance images and two dimensional (2D) portrait images and identified five genetic loci associated with certain facial phenotypes [102]. The lack of SNPs with strong effects, and the low degree of variance explained, indicate that facial phenotypes are likely to be highly polygenic and that larger sample sizes and more refined methods are needed to increase the variance explained for these phenotypes.

In this work, we correlated polygenic variation to specific 2D facial characteristics across a population of 1,338 individuals. Our aim was to prove that genetic variants can be used to construct portrait sketches of individuals; an overall approach we will refer to as appearnetics. The focus is on finding a one-way mapping between a genome-wide statistical SNP model and a facial statistical shape model. First, the statistical shape model, denoted eigenshapes, is constructed by principal component analysis (PCA). Then, the statistical SNP model, denoted eigengenes, is constructed by supervised PCA, where SNPs likely to be associated with the eigenshapes are included in the model. Finally, a supervised one-way mapping between eigengenes and eigenshapes is performed and evaluated by classifying individuals to simple facial archetypes in a cross-validation setup on an unseen data set.

Understanding the exact genetic variation responsible for specific facial characteristics is a colossal task. In this study we correlate a number of genetic variants to facial features and attempt to further predict these features. To our knowledge this is the first time this has been attempted.

Materials and Methods

Ethics Statement

All biological samples were collected after approval had been obtained from the National Bioethics Committee and the Icelandic Data Protection Authority. Written informed consent was provided by the participants. In order to protect personal information on health and disease all samples were encrypted through a third independent party in accordance with the regulations of the Icelandic Data Protection Committee [72].

Overview

This study focuses on finding a one-way mapping between a genome-wide SNP model (eigengenes) and a facial shape model (eigenshapes). For this, millions of SNPs were available, therefore a modified version of supervised principal components [5] was used to ensure that the eigengenes model was built with SNPs correlated to the eigenshapes. The method contained the following steps.

1. Data acquisition and preprocessing
   (a) Ascertainment
   (b) Genotyping
   (c) Image acquisition and normalization
   (d) Facial landmarks annotation
2. Model construction
   (a) Eigenshapes via PCA
   (b) Eigengenes via Supervised PCA
   (c) Supervised one-way mapping between eigengenes and eigenshapes

In the following the individual steps are described in detail.

Ascertainment

The ascertainment of Icelandic participants has been described in detail elsewhere [142]. Of these 1,338 were acquired with facial image data, 72 participants were removed due to non-Icelandic ethnicity, non-neutral facial expression, or non-frontal face direction to the camera, resulting in a remainder of 1,266 participants. All participants were adults (aged 18 or older).

Genotyping

Participants were genotyped using IlluminaHumanHap and IlluminaOmniExpress arrays and long range phased for efficient imputation of markers as described earlier [84]. The genotyping was performed at deCODE genetics in Reykjavik, Iceland.

Image acquisition and normalization

Frontal facial images were recorded in a controlled environment at enCODE and Landspitali University Hospital in Reykjavik, Iceland. At both sites identical photo studios were set up and all images were taken with a Canon IXUS 95, 10 megapixels camera, mounted on a custom-built rig. All participants were recorded sitting on a height-adjustable chair, with a horizontal and vertical ruler in the image frame to determine the scale. Images were corrected for distortions, mainly introduced by the camera lens, by using camera calibration [170].

Facial Landmarks

After the image acquisition, the frontal faces were annotated by in-house developed image analysis software imploring a semi-automatic annotation scheme. First, face and eyes were automatically detected using the Viola-Jones object detection algorithm [158], and then an Active Appearance Model placed the remaining annotation points [34]. The annotation scheme consisted of 73 anatomical landmarks and pseudo-landmarks and was similar to the one used in [50], an example is shown in Figure E.1. Finally, the annotation was revised and adjusted manually by a trained operator.

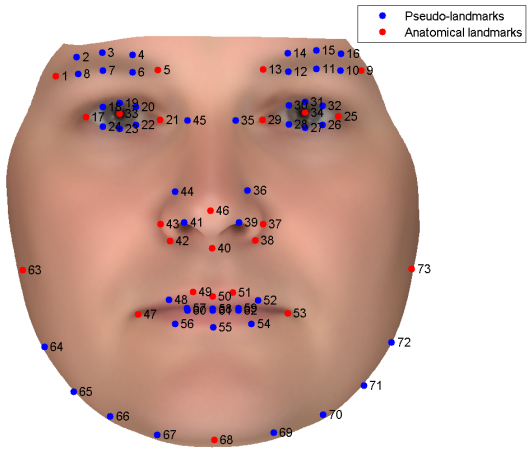


Figure E.1: 73 points annotation scheme.

Eigenshapes via PCA

To reduce the effect of affine transformations due to rotation, scaling and translation between the facial landmarks of different participants a Procrustes Analysis (PA) [69] was performed. This resulted in aligned shapes with affined transformations removed. The mean shape was subtracted from the aligned shapes and the statistical shape model (eigenshapes) was constructed by applying a Principal Components Analysis (PCA). The resulting model accounts for 95% of the variance of the original data, filtering out high-frequency variation present in the last 5%. The participants can now be described by the eigenshapes

$$X = \mu + b\Phi + \epsilon, \tag{E.1}$$

where X is the original face shape, μ the mean face shape, b the principal component scores for the participants, Φ the eigenshapes and ϵ the residual error. For the supervised PCA, explained in the next section, b, adjusted for year of birth and gender, was used as the phenotype.

Eigengenes via Supervised PCA

A genome-wide analysis was performed on two-thirds of the data for each of the eigenshapes to select the SNPs most correlated with the eigenshapes. Genome-wide significant p-values (<10-8) for single SNPs were not expected; instead we focused on multiple SNPs to explain a given eigenshape by retaining SNPs with association p-values less than 0.005, yielding $\sim 95,000$ SNPs per eigenshape. The genotypes were coded to an additive model and a PCA was applied to construct a genotype model (eigengenes) corresponding to each eigenshape. As previously, the model was confined to explain 95% of the variance. From this the eigengenes of a participant can be encoded to the very compact principal components scores, similar to equation 1. Unknown samples can now be encoded using the eigenshapes and eigengenes models.

Supervised one-way mapping between eigengenes and eigenshapes

The remaining one-third of the data, not used previously, were used to build the predictive models. Linear regression was implemented in a 40-fold cross-validation scheme, dividing the data into training, calibration and test-sets, to learn the transformation from the individual eigengenes to the corresponding eigenshape. In each iteration the mapping was estimated on 95% of the data, 2.5% was used for the calibration of the model and 2.5% was used for testing the performance of the model.

To overcome the problem of having much fewer participants than independent variables (eigengenes), which leads to over-fitting the model to the data, the eigengenes were sorted by their correlation to the corresponding eigenshape and a subset was selected. This subset was used to predict a given eigenshape in a linear regression. The size of the subset was determined by testing the prediction performance of the model on the calibration set so that

$$\underset{\theta}{\operatorname{argmax}}\ corr(\bar{y}_\theta, y)\ s.t.\ \theta \in \{1, \ldots, n\}, \tag{E.2}$$

where $\bar{y}_\theta$ is the predicted values, y the true values and n is the number of eigengenes. The resulting model was then applied to the test set and the performance was evaluated by comparing the predicted values to the target values. The supervised one-way mapping between eigengenes and eigenshapes was performed

five times with different random folds initializations with an equal gender distribution when applicable, to report the sensitivity of the results in standard deviations.

Results

In this proof-of-concept study we perform two evaluations on the proposed method. First we test whether it is possible to find a one-way mapping between the eigengenes and a single eigenshape. Secondly we classify individuals to simple facial archetypes.

827 individuals, two-thirds of the data, were used to construct the eigenshapes and eigengenes models, accounting for 95% of the variance of the original image data. This resulted in a model consisting of 37 eigenshapes and ~ 670 eigengenes per eigenshape. The remaining 439 individuals were used to predict the one-way mapping between eigengenes and eigenshapes in a 40-fold cross-validation scheme. The analyses were conducted five times to report the sensitivity of the results in standard deviations. Furthermore, the analyses were also conducted on men and women separately.

Correlation of predicted and true eigenshape scores

We build a linear regression model using the eigengenes with the highest correlation to the eigenshape as predictors to test whether we can predict eigenshape scores significantly correlated with the original scores. We found eigenshapes 3, 27 and 11 for the groups men & women, men and women, respectively, predicted with the best performance (Table E.1). Since it was our goal to identify the best fit between ~ 670 eigengenes, we corrected for multiple testing (Bonferroni correction), lowering the significance threshold from 0.05 to 7.5e-5. As expected, the prediction performance varied for individual eigenshapes because some eigenshapes encode non-facial variation incurred during photo acquisition. For example, eigenshapes 1 and 5 mainly encode head rotation in two directions, yaw and pitch, respectively.

Table E.1: Best performing eigenshapes

Best performing eigenshapes (men & women, n=439)		
Eigenshape number	**Pearson correlation**	**p-value**
3	0.2332 ± 0.0225	**7.81e-07**
7	0.1857 ± 0.0094	9.11e-05
17	0.1799 ± 0.0135	1.50e-04
34	0.1777 ± 0.0262	1.82e-04
8	0.1774 ± 0.0039	1.87e-04

Best performing eigenshapes (men, n=231)		
Eigenshape number	**Pearson correlation**	**p-value**
27	0.3114 ± 0.0195	**1.38e-06**
34	0.2476 ± 0.0152	1.44e-04
30	0.2402 ± 0.0208	2.29e-04
37	0.2158 ± 0.0153	9.66e-04
15	0.2051 ± 0.0234	1.72e-03

Best performing eigenshapes (women, n=208)		
Eigenshape number	**Pearson correlation**	**p-value**
11	0.3015 ± 0.0049	**9.65e-06**
28	0.2767 ± 0.0218	**5.21e-05**
3	0.2725 ± 0.0229	**6.83e-05**
16	0.2420 ± 0.0336	4.29e-04
12	0.2364 ± 0.0434	5.86e-04

Best performing eigenshapes for the three groups men & women, men and women. Bold signifies a statistical significant correlation (Bonferroni corrected).

Classifying persons to simple facial archetypes

The one-way mapping in the previous section revealed that, in this proof-of-concept study, we are missing power to predict many of the eigenshapes. Consequently we chose to classify to simple facial archetypes. The test was conducted in a cross-validation setup, where the calibration set was used to detect the eigenshapes predicted with the highest performance. We classified into two facial archetypes using Mahalanobis' distance. We make the facial archetypes by choosing the best performing eigenshape from the calibration set and constructing two archetypes with a distance of -2 and +2 standard deviations away from the mean along the eigenshape direction. We constructed two additional

archetypes by taking the second best performing eigenshape, yielding a total of four facial archetypes. We add two extra archetypes in a similar fashion to obtain 6, 8 and 10 facial archetypes. We found that the simple archetype classification performed better than random for the group of combined men & women (Figure E.2). Classification on both genders separately resulted also in similar predictions (data not shown). An illustration of the six best performing archetypes (three eigenshapes) for the three groups is shown in Figure E.3.

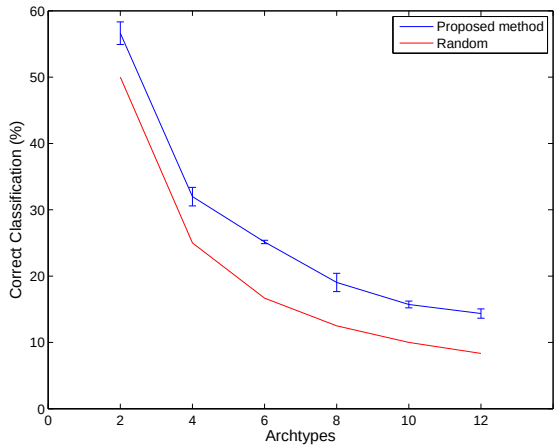


Figure E.2: Archetype classification performance for the combined group of men & women.

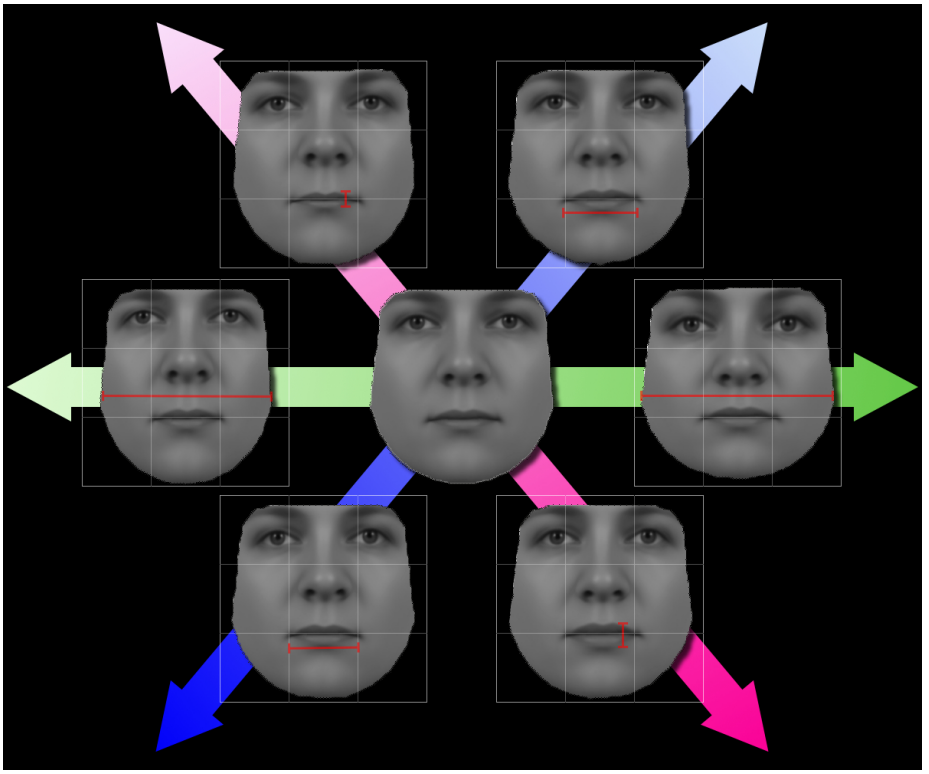


Figure E.3: The six best performing archetypes (three eigenshapes) for the three groups men & women (green - eigenshape 3), men (blue - eigenshape 27) and women (magenta - eigenshape 11).

Discussion

This work demonstrates the potential to predict complex personal facial characteristics from genetic information only. Despite the modest number of indviduals and high polygenicity, we show that several complex facial characteristics can be partially predicted. As shown in Paternoster *et al.* [119], the height-to-width ratio of the face is the feature showing the highest potential in the context of prediction of appearance based on genetics. In addition to this our model predicts fullness of female lips, and a slight variation in mouth width in men.

Prediction of specific facial characteristics can be accomplished without under-

standing the exact genetic variation responsible for these. In this study we prove the validity of the concept, however, large sample sizes, better image quality, and better models are needed to determine to which extent facial characteristics can be revealed from DNA only. Next steps should go beyond shape characteristics by including texture information (texture and color of hair, eye, skin etc.) into the models.

A potential caveat in this study is the assumed independence between cross-validation folds. If this criterion is not fulfilled then the results will be biased towards that dependence. Therefore one concern of this study is the use of only Icelandic individuals who as a population have a larger genetic similarity than other populations.

Another important aspect is the protection of the anonymity of DNA donors, which has been an issue for years [42]. Though this study shows the possibility that appearnetics at some point may succeed in predicting portraits resembling the real face, each prediction will still only be a resemblance. Such an image may be used for screening suspects for specific facial features, but it is very unlikely that it would be precise enough for actual identification. An image with a high likeness to a person could be a serious threat to anonymity, however appearnetics can only be a threat to anonymity to the extent that a person's appearance is similar to the predicted face, combined with knowledge about how similar the remaining people in a given cohort are to the predicted image.

The polygenic risk scores that are currently being calculated for a vast number of different phenotypic traits indicates that genotype data alone will only enable incomplete prediction of most traits [156, 166]. For appearnetics to succeed, finding the best suited prediction method for solving the mapping from DNA to facial characteristic is therefore only part of the obstacle; acquiring standardized facial characteristics (images) is also an important challenge. Epigenetic data is another feature that in the future may help predict facial characteristics [60]. Furthermore, even if only a small part of the epigenetic effects ever becomes measurable based on available DNA, appearnetics still has the potential to produce images giving valuable clues to the appearance of the person to whom the DNA belongs.

Acknowledgments

The research leading to these results has received support from EU funded FP7-People-2011-IAPP grant PsychDPC (GA 286213). T.H.P. was supported by The Danish Council for Independent Research Medical Sciences and the The

Alfred Benzon Foundation.

Appendix F

Discrete facial feature of NRNX1 deletions carriers

Thomas F. Hansen, Jens Fagertun, Linh Duong, Anders Rosengren, Erik Sørensen, Henrik Ullum, Rasmus Larsen, Rasmus R. Paulsen and Thomas Werge

Abstract

Genome wide association studies have identified a number of copy number variations (CNVs) that confer high risk of both psychiatric and other neurodevelopmental disorders. These findings have highlighted both the pleiotropic and heterogenic nature of the etiology underlying severe mental disorders. The objective is to examine whether there are, independently of affection status or diagnose categories, specific traits for carriers of such risk. We focused on facial features, as previous studies have found a general increase in minor physical anomalies among patients with schizophrenia and autism, in particular facial features. Images from two unrelated families with multiply affected by severe mental disorders with six carriers of deleterious mutations in neurexin 1, and 605 unrelated healthy blood donors were annotated and analysed using Mann-Whitney test. We found a nominal significant thicker upper lip, vermillion, and a smaller nose width of individuals carrying a deleterious mutation in neurexin 1 as compared with their relatives not carrying deleterious mutations in neurexin 1. Keywords: *Schizophrenia, Mental disorders, deleterious mutations in Neurexin 1, Neurexin 1, Facial traits, Facial analysis, Copy Number Variation, CNV, Minor physical anomalies.*

F.1 Introduction

The genome wide assessment of copy number variation (CNV) has led to identification a number of CNVs that confer risk of schizophrenia and other neurodevelopmental disorders, such as autism and mental retardation. These findings highlight both the pleiotropic and heterogenic nature of the etiology of severe mental disorders, e.g. both deletions on chromosome 1q21 and in neurexin 1 (NRXN1) have been associated with both schizophrenia [131, 143] and autism [67, 111] and has been reported in patients with mental retardation and recently in patients with epilepsy [168, 46, 45, 32, 66]. An important question is whether specific clinical or biological features of such risk factors are manifest in carriers across diagnostic categories (i.e. independently of affection status).

Previous studies have found an general increase in minor physical anomalies among patients with schizophrenia and autism, in particular facial features [134, 162]. However, no consensus on disorder-specific traits has emerged from these studies, which has traditionally been ascribed to the lack of standardized, objective measures and an underlying etiological heterogeneity of these disorders. There are some case reports showing gross facial features among carriers of deletions on chromosome 1q21.1, 15q11.2, 15q13.3 and 22q11.2, all associated with schizophrenia, autism and mental retardation [143, 44, 73, 21, 139, 146].

We have previously reported on deleterious neurexin-1 mutations in two families multiply affected with severe mental disorders [46, 45]. Here, we have used detailed facial images of affected and non-affected family members with and without neurexin-1 mutations as well as of unrelated, healthy controls to identify a neurexin-1 specific facial trait.

F.2 Methods

F.2.1 Participants

The study included 13 members of two unrelated families multiply affected by severe mental disorders (see Figure F.1), 6 of which carried one of a total of three different deleterious mutations in NRXN1. In brief, the two families - described in detail in two recent publications [46, 45] - included subjects with bipolar disorder, schizophrenia, epilepsy and mental retardation. A member of family 2 carrying mutations in both NRNX1 genes presented with hypertonic spasm [46] impeding the acquisition of a usable facial images. Also, the study

included 605 healthy and unrelated subjects recruited among unpaid, volunteers of the Danish Blood Donor Study (www.DBDS.dk; [122]).

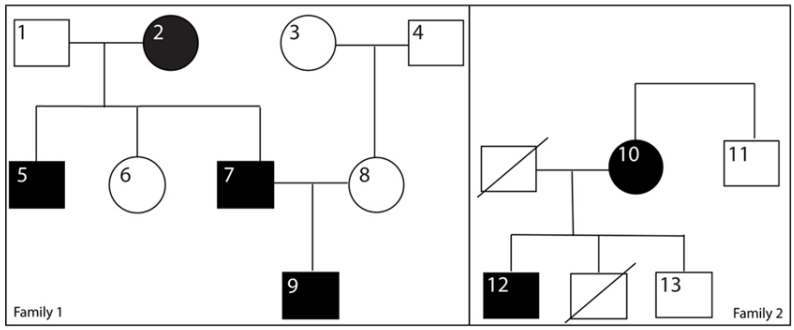


Figure F.1: The pedigrees of the two families assessed. Circles are females, squares are males, black filled shapes are NRXN1 deletion carriers and crossed shapes are deceased.

F.2.2 Facial Images

All individuals had a fixed-sized sticker placed on their forehead in order to compare absolute measures. Two-dimensional, 2D, images of the family members were captured using a Canon EOS 500D camera. Three-dimensional, 3D, of unrelated healthy controls subjects using a Canfield Vectra M3 Imaging System, permanent DBDS facility at Glostrup University Hospital.

F.2.3 Face Annotation

Seventy-four distinct facial points were manually annotated, blinded for the carrier status of participants. Three operators, supervised by JF, did the annotation the pictures. The annotation scheme captures anatomical features, such as eyes, mouth, and jaw. The first seventy-three facial points are described in by Fagertun *et al.* [58, 50], the additional point is used to describe the middle

hairline. These annotation points were then used to measure the 13 vertical and the 13 horizontal distances of the face as illustrated in Table F.1 and Figure F.2. All facial distances were normalized to the centroid size of the face, calculated as the average distance between the centre of gravity of all annotation points and the annotation points, respectively. This allows direct comparison between individuals. The fixed-sized sticker was used to estimate absolute facial distances in Euclidian space. The 3D facial images of the 605 unrelated controls were projected into a 2D plane, simulating a standard 2D image, and annotated in concordance with that of the 2D images.

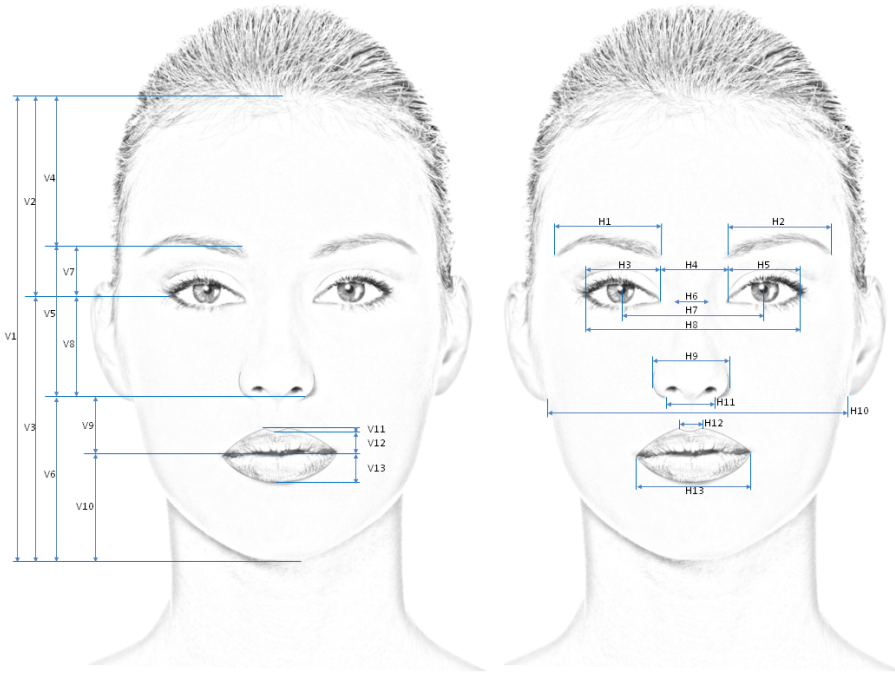


Figure F.2: Face distances illustrated.

F.2.4 Statistical approach

Non-parametric statistical analysis was used (Wilcoxon/Mann-Whitney), was used in MATLAB version 2010b.

Distance	Name	Description
H1	Width of the right eyebrow	Distance between the right eyebrow lateral and medial points
H2	Width of the left eyebrow	Distance between the left eyebrow lateral and medial points
H3	Width of the right eye	Distance between the right eyes medial and lateral canthus
H4	Inner distance between eyes	Distance between the left and right eyes medial canthus
H5	Width of the left eye	Distance between the left eyes medial and lateral canthus
H6	Width of the nose root	Distance of the nose root, at the junction of the eyes medial canthus
H7	Center distance between eyes	Distance between the left and right pupil centers
H8	Outer distance between eyes	Distance between the left and right lateral canthus of the eyes
H9	Width of nose tip (outer)	Distance between the outer left and right alar-facial groove
H10	Width of jaw	Distance between ear attachments
H11	Width of nose tip (inner)	Distance between the inner left and right alar-labial groove
H12	Width of the philtrum columns	Distance between the philtrum columns where they meet the vermillion border (white role)
H13	Width of the mouth	Distance between the oral commissures
V1	Face height	Distance between the lowest point of central hairline at the forehead and the lowest point of the central chin
V2	Upper face height	Distance between the average point of the eyes medial and lateral canthus and the lowest point of central hairline at the forehead
V3	Lower face height	Distance between the average point of the eyes medial and lateral canthus and the lowest point of the central chin
V4	Height of forehead	Distance between the average point of the left and right eyebrow lateral and medial points and the lowest point of central hairline at the forehead
V5	Distance from eyebrows to the lower nose	Distance between the average point of the left and right eyebrow lateral and medial points and the columella-labial junction
V6	Distance from the lower nose to the chin	Distance between the columella-labial junction and the lowest point of central chin
V7	Distance from the eyebrow to the eyes	Distance between the average point of the eyes medial and lateral canthus and the average point of the left and right eyebrow lateral and medial points.
V8	Height of nose	Distance between the average point of the eyes medial and lateral canthus and the columella-labial junction
V9	Distance from the lower nose to the mouth	Distance between the upper lips lower border and the columella-labial junction
V10	Distance from the mouth to the chin	Distance between the lower lip upper border and the lowest point at the central chin
V11	Height of the center of the cupid's bow	Distance between the lowest point of the philtrum dimple to that of the highest point of the column.
V12	Upper lip thickness	Distance between the lowest point of the cupid's bow to the center of the visible internal border of the upper lip
V13	Lower lip thickness	Distance between the point at the center of the lower vermillion border (white roll) to the center of the visible internal border of the lower lip

Table F.1: Description of face distances.

F.3 Results

Initially, the two families were collapsed and the 26 normalized, horizontal and vertical distances were compared between carriers and non-carriers of neurexon-1 deleterious mutations, using Mann-Whitney test. A nominal association between the both a smaller nose root (distance H6, p=0.0051) and a larger upper lip thickness (distance V12, p=0.030) and NRXN1 deletions carriers, see Figure F.3A.

To examine whether the observed associations were due to facial characteristics serendipitously present in the two families (a likely risk as two mutation carriers only represent one family), we analyzed all members of the two families against 605 unrelated, healthy subjects. These analyses confirmed that none of the two distances were significantly associated with family membership (p>0.42).

Post-hoc analysis of the distances V12 and H6, respectively, was done using absolute distances, see methods. We found an association between V12 and NRXN1 carriers as compared with non-carriers in the collapsed families, p=0.030, see Figure F.3B . However, we did not see any association to H6. Again, to exclude the association as serendipitous findings we analyses the two families against 605 unrelated, healthy subjects. We find no association with V12 and family membership, p=0.97.

F.4 Discussion

To our knowledge, this is the first study to show a discrete facial trait to be associated with deleterious NRXN1 mutations. Besides identifying a putative facial marker for NRXN1 deletion carriers, it also supports the notion that there may be common features between carriers of high-risk genetic factors with different disease entities.

This study is to be considered an explorative study, as it comprises a small effect size and does not include replication of the findings in an independent sample. We find nominal significant association between deleterious NRXN1 mutations and facial traits, where the most significant findings were only resistant for a correction factor of 22 (Bonferroni). We cannot exclude that for some of the annotation points, the age and BMI may affect the annotation. Still there are many indications for a true finding: 1) The association was present when analyzing two independent families and different generations, and 2) does not

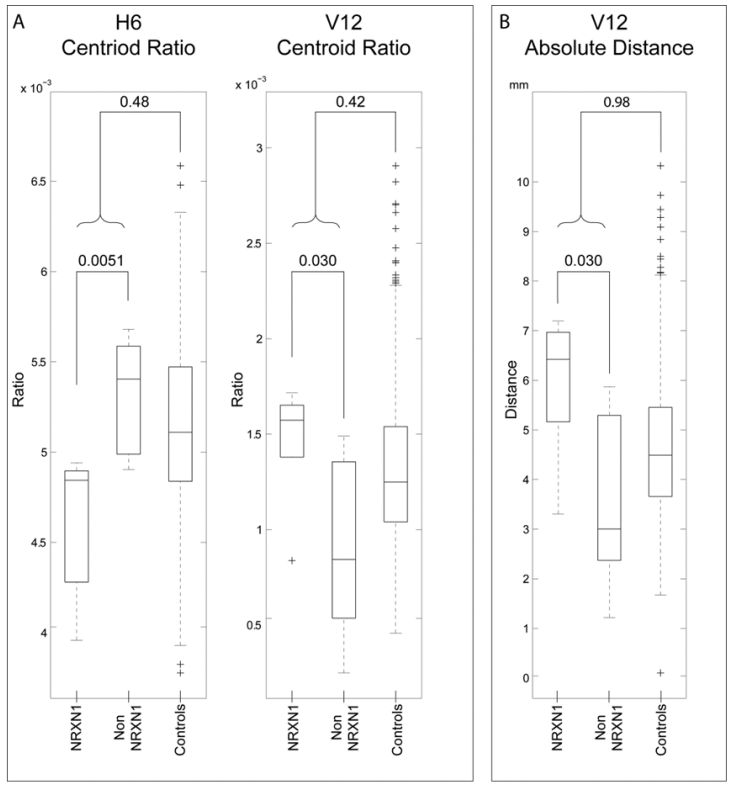


Figure F.3: Box-plot. Facial distances nominal associated with NRXN1.

association between family-membership. 3) The association of thickness of the upper lip was also present when analyzing the absolute distances, which indicates that it arises from the facial trait per se. All in favor of true association between the deleterious NRXN1 mutation and the thickness of the upper lip.

As described, deletions on chromosome 22q11.2 giving rise to diGeorge syndrome also leads to highly elevated risk of developing schizophrenia and autism [143, 146]. Carriers of these 22q11 deletions have characteristic both subtle or gross dysmorphic features. Among the most characteristic features are subtle malformations of the mouth and nose and more grossly malformations as cleft lip/palate and velopharyngeal insufficiency [73, 117]. Thus, both deleterious mutations in NRXN1 and chromosome 22q11 deletions elevated the risk of developing schizophrenia and autism. Further, considering the embryonic development, they also affect fusion process between the maxillary and nasomedial processes giving rise to malformations of the lip. The question is whether the resemblance between the deleterious mutations arises because of some common molecular pathology. Thus, the mutations affect the same pathway that early in development gives rise to erroneous fusion process between the maxillary and nasomedial processes but at the same time or later disturb the neurodelveoplement resulting in a increase risk of developing psychiatric disorders. Alternatively, the effect is merely an indication of a disturbed neural development during 5-8 gestational-week of the fetus developmental, i.e. the gestational weeks of facial formation.

In perspective, the putative association between facial traits and increased risk of developing a psychiatric disorder, may seed development of non-invasive tools that can assist assessment of patients either to direct treatment or to support diagnosing.

F.5 Acknowledgement

We thank the two families for participating. We acknowledge the expert assistance of numerous mental health professionals in the various clinical departments. In particular, we appreciate the hospitality given by the staff at Danish Blood donor Bank, Glostrup Hospital.

This project was sponsored by grant to Thomas Werge from The Danish National Advanced Technology foundation (Jnr. 001-2009-2) and Lundbeck Foundation (R34 A3243) and by grants to Thomas Hansen from the Danish Main Capital Research Foundation of Health. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the

manuscript. The authors have declared that no competing interests exist.

Appendix G

Detection of minor physical anomalies in the face associated with severe schizophrenic disorders

Jens Fagertun, Christian K. Møller, Thomas Werge, Rasmus R. Paulsen and Thomas F. Hansen

Abstract

This paper presents an objective analysis of the connection between minor physical anomalies in the face and mental disorders within the schizophrenic spectrum. We use high resolution 3D scans to associate human understandable distances to the disorder. In a classification experiment we show very good performance in detecting patients from controls. Finally, we construct a "face map" that shows regions that associate with the schizophrenic disorder spectrum. Keywords: *Schizophrenia, Mental disorders, Facial traits, Facial analysis, Minor physical anomalies.*

G.1 Introduction

Down's, Klinefelter's , di George's and XXY syndromes all have very distinct anthropometric facial traits. Down's and di George's syndrome are associated with an increased risk of developing schizophrenia as well as other psychiatric disorders. Prior studies have suggested a connection between minor physical anomalies (MPA) and psychiatric disorders especially within the schizophrenic spectrum, where the MPA were located in the face and on the hands in particular. In Weinberg *et al.* [162] an overview of 13 studies from 1982 to 2005 is given that documents this connection.

Despite these previous studies the connection between MPA and schizophrenic disorders is not well defined and no underlying correlation has been established. An explanation could be that the presence of MPA is based on the judgment of a human observer and hence subjective, making it hard to find a correlation. In this work we address this problem by introducing objective measurements and analysis. We find a strong association between patients diagnosed with a disease in the schizophrenic disorder spectrum and distances and ratios in the human face. Furthermore, we map the human face for regions (MPA) that associate with the schizophrenic disorder spectrum.

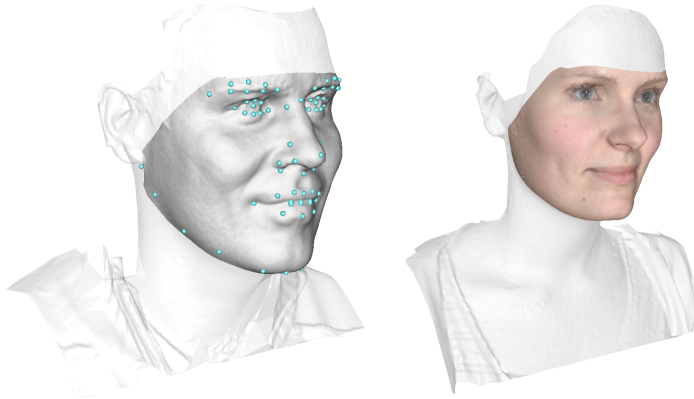
G.2 Methods

G.2.1 Data

The data used in this work consists of 1276 high resolution 3D facial scans divided between patients and controls. The control cohort is of 1061 healthy and unrelated subjects recruited among unpaid volunteers of the Danish Blood Donor Study [122]. The patients cohort consist of 215 individuals recruited at two psychiatric institutions in Denmark. All patients have one or more diagnoses within the following schizophrenic disorder spectrum: Paranoid schizophrenia, Hebephrenic schizophrenia, Catatonic schizophrenia, Undifferentiated schizophrenia, Residual schizophrenia, Simple schizophrenia and Schizophrenia unspecified.

All facial scans were captured using a Canfeld Vectra M3 Imaging System. Each 3D facial scan contains about 70,000 to 100,000 3D points and has shape information (x-, y-, z-point positions) and texture information (red, green, blue intensities) for every 3D point. Every scan was manually annotated by seventy-

four distinct facial landmarks, blinded for patient/control status of participants. The annotation scheme captures anatomical features such as eyes, mouth, and jaw. The first seventy-three facial points are described by Fagertun *et al.* [50], the additional point is used to describe the nasion point, the midline point just superior to the nasal root overlying the naso-frontal structure. The seventy-four facial landmarks are shown in Figure G.1.



(A) (B)

Figure G.1: Annotation landmarks (A) and dense correspondence mesh (B).

G.2.2 Data sets

The 3D scans and annotations were used to create three data sets for use in this work:

- *Thirty-four human understandable facial distances.*
  Thirty-four human understandable facial distances were constructed from the landmarks. The distances are shown in Figure G.2 and capture facial distances such as "width of the mouth", "length of the nose", etc.
- *The original seventy-four landmarks.*
  Used to analyze basic facial shape variation.
- *A dense point to point correspondence mesh.*
  To analyze facial shape variation at positions not annotated by landmarks, a dense point correspondence is created [121]. The dense correspondence mesh contains 39653 points, generated from the original scans and the

seventy-four landmarks. An example of the dense correspondence mesh is shown in Figure G.1.

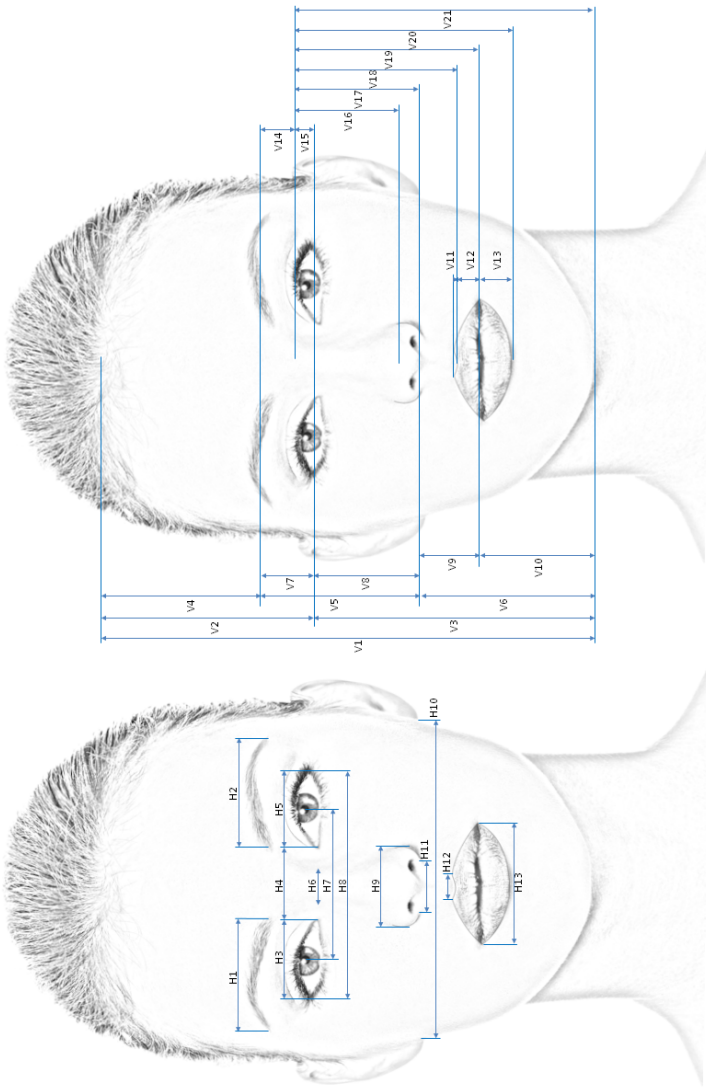


Figure G.2: Human Understandable Facial Distances.

G.2.3 Analysis

All human understandable facial distances were normalized to the centroid size of the face, calculated as the average distance between the center of gravity of all annotation points and the annotation points. This allows direct comparison between individuals. We then perform a non-parametric Wilcoxon rank sum test to investigate if the patients are significantly different from the controls.

Based on the two correspondence data sets a classification experiment is performed. We align the correspondence data sets by Procrustes Analysis (PA) [69], to remove variation due to rotation, scaling and translation. We then make a feature extracion/dimmensionallty reduction by principal component analysis (PCA). Finally we classify by linear discriminant analysis (LDA) to the two class problem, patient or control. The analysis is performed in a double five folds cross-validation setup with the following pseudo code:

```
Align all data via a PA;
Divide data into five folds;
for every fold do
    Construct a PCA of one fold, retaining 100% of the variance;
    Project the remaining four folds into this;
    for every of the remaining 4 folds not used for PCA do
        Train a LDA on 3 of the 4 folds;
        Test performance on the last fold;
    end
end
```

Algorithm 1: Double five folds cross-validation setup

Finally, the dense correspondence data set is used to construct a "Face Map" over MPA in relation with the schizophrenic disorder spectrum. In this way we visualize the importance of the different human facial features when detecting schizophrenic disorders. We do this by a multivariate equivalent of the two-sample t-test, the Hotelling's two-sample T-squared statistic

$$T^2 = \frac{n_x n_y}{n_x + n_y}(\overline{x} - \overline{y})^T S^{-1}(\overline{x} - \overline{y}), \tag{G.1}$$

where $\overline{x}$ and $\overline{y}$ are the means, and n_x and n_y are the sizes of both samples

respectively and the unbiased pooled covariance matrix estimate is

$$S = \frac{\sum_{i=1}^{n_x}(x_i - \overline{x})(x_i - \overline{x})^T + \sum_{i=1}^{n_y}(y_i - \overline{y})(y_i - \overline{y})^T}{n_x + n_y - 2}. \tag{G.2}$$

The Hotelling's two-sample T-squared statistic can, for the three-dimensional case, be related to the F-distribution by

$$\frac{n_x + n_y - 4}{(n_x + n_y - 2)3} T^2, \tag{G.3}$$

and a p-value for statistical significance can be obtained.

G.3 Results

The Wilcoxon rank sum test is performed on the human understandable facial distances. Using a conservative Bonferroni correction for multiple testing on the thirty-four facial distances lowers the significance threshold from 0.05 to 0.0015. The statistical significant facial distances are shown in Table G.1. Clearly some facial distances are significant, where the width of the mouth (H13) shows to be the most predominant, with very low p-value.

Performing the classification experiments on both the original seventy-four landmarks and the dense correspondence mesh data sets in the double five folds cross-validation setup showed very good performance (>87% correct classification). Results are presented in Table G.2.

Finally, a Schizophrenia Disorders Face Map is presented by the Hotelling's two-sample T-squared statistic and Wilcoxon rank sum test. Figure G.3 shows which points on the face are significant (p-value from Hotelling's T^2) for detecting schizophrenia disorders. Figure G.4 shows only Bonferroni significant relation areas (Wilcoxon) for facial width, height and depth (x-, y-, z-plans), respectively. A Bonferroni correction for multiple testing will lower the significance threshold of 0.05 to 1.26e-06 when doing 39,653 tests. In this case the Bonferroni correction must be considered a conservative estimate as the different tests must be considered very correlated.

Facial distance	Wilcoxon rank sum test
H13	3.52e-35
V17	3.00e-11
V16	5.74e-10
H11	7.03e-10
V18	1.52e-08
V12	1.67e-08
V13	7.23e-07
V19	2.79e-06
H4	1.13e-05
H7	1.41e-05
H3	2.17e-04
H5	2.40e-04
H2	9.09e-04
V11	9.61e-04

Table G.1: Facial Distances Associated with schizophrenic disorders.

Data set	Control group	Patient group	Total
Seventy-four annotation landmarks	89.4%	74.0%	86.8%
Dense correspondence mesh	89.3%	79.7%	87.4%

Table G.2: Correct Classification of Schizophrenia Disorders.

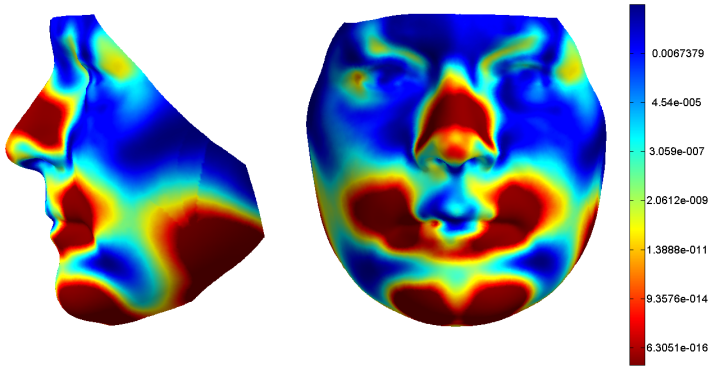


Figure G.3: Facial morphology relation with schizophrenic disorders (p-value from Hotelling's T^2).

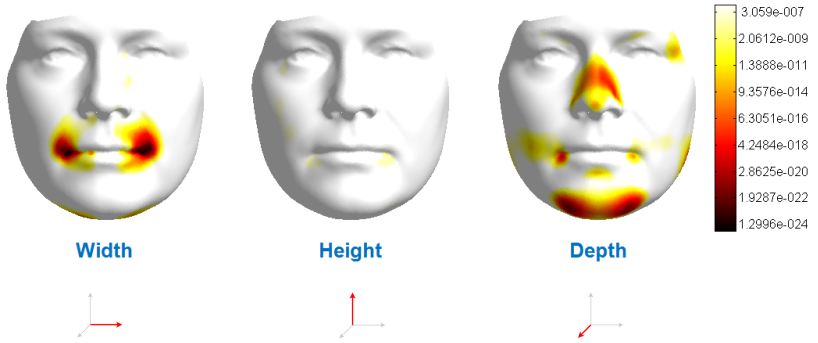


Figure G.4: Bonferroni significant areas with respect to schizophrenic disorders (Wilcoxon) for facial width, height and depth (x-, y-, z-plans), respectively.

G.4 Discussion and Conclusion

In this work we have shown that some small distinct facial morphology changes (MPA) seem to associate with schizophrenic disorders, where the width of the mouth shows to be the most predominant with a p-value of 3.52e-35. In this context it is good to see that width of the face has a high p-value > 0.5 (Bonferroni corrected), because the body mass index (BMI) also correlates with schizophrenic disorders. The very high performance of the LDA classifier that correctly classified 87.4% of the individuals into patient/control classes suggests that looking for MPA in high accuracy 3D scans can give a risk indication that could assist doctors in the diagnostic task. Finally, we presented a face map illustration of where in the face a MPA that is associated with schizophrenic disorders is likely to be located.

Here we should note that the individuals in this study consist of patients and blood donors. Blood donors are more healthy than the average population. The reported results on performance will probably be decreased if exposed to data that include samples from the general public.

Appendix H

Elastic Appearance Models

Mads F. Hansen, Jens Fagertun and Rasmus Larsen

Abstract

This paper presents a fusion of the active appearance model (AAM) and the Riemannian elasticity framework which yields a non-linear shape model and a linear texture model – the active elastic appearance model (EAM). The non-linear elasticity shape model is more flexible than the usual linear subspace model, and it is therefore able to capture more complex shape variations. Local rotation and translation invariance are the primary explanation for the additional flexibility. In addition, we introduce global scale invariance into the Riemannian elasticity framework which together with the local translation and rotation invariances eliminate the need for separate pose estimation. The new approach was tested against AAM in three experiments; face labeling, face labeling with poor initialization and corpus callosum segmentation. In all the examples the EAM performed significantly better than AAM. Our Matlab implementation can be forwarded on request.

H.1 Introduction

Generative model-based image interpretation for image segmentation, registration and labeling plays a central role in many image analysis applications. One of the most successful model-based approaches is the *Active Appearance Model*

by Cootes *et al.* [33] which applies learned shape and texture (intensity) priors from a training set. The method has been applied to medical image segmentation [154, 113, 10], eye tracking [75, 99], object tracking [144] and facial recognition [47, 2, 71] with success.

AAM creates a shape model from set of training shapes by modeling them in a linear subspace. Linear generative shape models will in general not be able to capture complex shape variations as biological shapes rarely can be embedded in a linear subspace. This may result in folding shapes and poor registration/segmentation. Instead, we will assign an individual displacement to each point in the mean shape and construct a statistical local deformation prior which measures how well a given local deformation matches the local deformations in the training set. We use the Riemannian elasticity framework [123] to model this prior. Riemannian elasticity priors can capture complex deformations as they are locally rotation and translation invariant.

H.2 Previous work

Many improvements to AAM have been proposed to strengthen the approach with respect to efficiency, discrimination and robustness, see Gao *et al.* [63] for a complete survey. Many improvements deal with reducing the computational cost by tuning the optimization, e.g. the inverse compositional algorithm by Baker *et al.* [6], or by compressing the texture model, e.g. the wavelet and wedgelet model by Larsen *et al.* [98]. Other research groups have increased the robustness of AAM by combining it with the active shape model [150] or by making it light invariant [91]. Suggestions to improve the accuracy of the AAM include the maximum a posteriori formulation by Cootes *et al.* [35] and inclusion of local image structure by Scott *et al.* [136].

To our knowledge Christoudias and Darrell [30] are the only researchers to propose non-linear shape modeling extensions to AAM. There are complementary methods to AAM who use non-linear shape modeling, e.g the deformable M-rep [125].

H.3 Linear appearance modeling

In AAM the goal is to construct a compact parametrization of the variability of an object, described by its shape and texture, from a set of (semi-)manually

labeled images. Generalized Procrustes analysis filters out rotation, translation and scaling effects from the landmarks yielding a mean shape and a set of aligned shapes. A linear shape model follows from a truncated matrix decomposition, e.g. Principal Components Analysis, on the aligned shapes, i.e

$$\boldsymbol{v} = \bar{\boldsymbol{v}} + \boldsymbol{U}_v \boldsymbol{p}_v, \tag{H.1}$$

where $\bar{\boldsymbol{v}}$ is the mean shape and $\boldsymbol{U}_v$ contains the modes of shape variation. Warping the textures to a common reference space, defined by the mean shape, creates shape-free textures, and another truncated matrix decomposition generates a linear texture model, i.e.

$$\boldsymbol{g} = \bar{\boldsymbol{g}} + \boldsymbol{U}_g \boldsymbol{p}_g. \tag{H.2}$$

A final truncated matrix decomposition on the combined set of shape and texture parameters yields a combined appearance model

$$\begin{bmatrix} \boldsymbol{v} \\ \boldsymbol{g} \end{bmatrix} = \begin{bmatrix} \bar{\boldsymbol{v}} \\ \bar{\boldsymbol{g}} \end{bmatrix} + \begin{bmatrix} w\boldsymbol{U}_v & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{U}_g \end{bmatrix} \boldsymbol{U}\boldsymbol{p}, \tag{H.3}$$

where w weights the relative importance between shape and texture.

H.4 Non-linear shape modeling using elasticity

Linear generative shape models, Equation H.1 and Equation H.3, will in general not be able to capture complex shape variations. This section introduces non-linear strain modeling and describes how a Riemannian elasticity prior can be constructed.

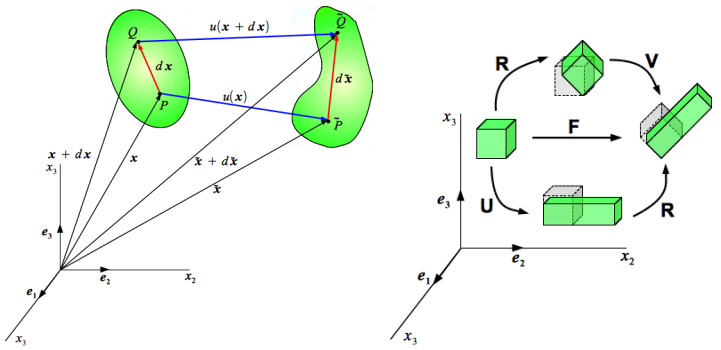
H.4.1 Representing local deformation

Consider the particle P and the neighboring particle Q in an elastic body positioned at $\boldsymbol{x}$ and $\boldsymbol{x} + d\boldsymbol{x}$, respectively (see Figure H.1(a)). The elastic body undergoes a diffeomorphic deformation, $\phi(\boldsymbol{x}) = \boldsymbol{x} + \boldsymbol{u}(\boldsymbol{x})$, such that the particle P and Q move to positions $\widetilde{P}$ and $\widetilde{Q}$ given by $\widetilde{\boldsymbol{x}}$ and $\widetilde{\boldsymbol{x}} + d\widetilde{\boldsymbol{x}}$. Obviously, there exists a linear mapping between $d\boldsymbol{x}$ and $d\widetilde{\boldsymbol{x}}$, i.e. $d\widetilde{\boldsymbol{x}} = \boldsymbol{M} d\boldsymbol{x}$, which represents the local transformation. From Figure H.1(a) it can be derived that

$$d\widetilde{\boldsymbol{x}} = d\boldsymbol{x} + \boldsymbol{u}(\boldsymbol{x} + d\boldsymbol{x}) - \boldsymbol{u}(\boldsymbol{x}). \tag{H.4}$$

By performing a first-order Taylor expansion of $\boldsymbol{u}(\boldsymbol{x} + d\boldsymbol{x})$ around P we get

$$d\widetilde{\boldsymbol{x}} \approx d\boldsymbol{x} + \boldsymbol{u}(\boldsymbol{x}) + \nabla \boldsymbol{u}(\boldsymbol{x}) d\boldsymbol{x} - \boldsymbol{u}(\boldsymbol{x}) = (\boldsymbol{I} + \nabla \boldsymbol{u}(\boldsymbol{x})) d\boldsymbol{x} = \boldsymbol{F}(\boldsymbol{x}) d\boldsymbol{x} \tag{H.5}$$



(a) Elastic body deformation (b) Left and right stretch tensors

Figure H.1: Representing local deformation.

Thus, for a sufficiently small $d\boldsymbol{x}$ we get the relationship $\boldsymbol{M} \approx \nabla\phi(\boldsymbol{x}) = \boldsymbol{F}(\boldsymbol{x})$.

As the deformation gradient tensor $\boldsymbol{F}$ is based on the displacement gradient it is translation invariant but not rotation invariant. However, $\boldsymbol{F}$ can be decomposed into an orthonormal tensor (representing the rotation) and a positive definite symmetric tensor (representing the stretch), i.e.

$$\boldsymbol{F} = \boldsymbol{R}\boldsymbol{U} = \boldsymbol{V}\boldsymbol{R}. \tag{H.6}$$

$\boldsymbol{U}$ and $\boldsymbol{V}$ are denoted the *right stretch tensor* and the *left stretch tensor* respectively. $\boldsymbol{U}$ and $\boldsymbol{V}$ have identical eigenvalues or *principal stretches* but different eigenvectors. The decomposition is illustrated in Figure H.1(b).

From the right stretch tensor we define a sequence of rotation and translation invariant tensors which are known as the *Lagrangian strain tensors*

$$\boldsymbol{E}_m = \frac{1}{2m}(\boldsymbol{U}^{(2m)} - \boldsymbol{I}). \tag{H.7}$$

These tensors essentially model the deviation of the right stretch tensor from the identity. The special case $\boldsymbol{E}_0 = \log(\boldsymbol{U})$ is called the *Hencky strain tensor* which will be the basis for our non-linear shape model.

H.4.2 Statistical deformation modeling

Given the Lagrangian strain tensors it is possible to formulate a rotation and translation invariant density function which quantifies the energy stored in an infinitesimal neighborhood due to the deformation. The most intuitive way to

build a strain energy density function is just to use the magnitude, defined by a suitable matrix norm (e.g. the Frobenius norm), of a suitable strain tensor. To protect against folding we want an energy density function, which approaches infinity when the approaches towards the black-hole deformation. This implies that the at least one of the eigenvalues of the strain tensor should go to infinity when the determinant of the Jacobian of the deformation gradient approaches zero. The Hencky strain tensor, $\boldsymbol{E}_0$, fulfils this requirement. Pennec *et al.* [123] proposed the *statistical Riemannian elasticity energy*

$$W(\boldsymbol{E}_0) = \begin{cases} \frac{1}{4}\|\text{vect}(\boldsymbol{E}_0 - \bar{\boldsymbol{E}}_0)\|^2_{\boldsymbol{\Sigma}^{-1}} & \text{if } \det(\boldsymbol{F}) > 0 \\ \infty & \text{otherwise.} \end{cases} \tag{H.8}$$

where Σ and $\bar{\boldsymbol{E}}_0$ are calculated from a set of n previously observed deformations. Hence,

$$\bar{\boldsymbol{E}}_0 = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{E}_{0,i}, \text{ and } \Sigma = \frac{1}{n}\sum_{i=1}^{n} \text{vect}(\boldsymbol{E}_{0,i})\text{vect}(\boldsymbol{E}_{0,i})^T. \tag{H.9}$$

H.5 Material and Methods

H.5.1 Warping and discretization

The deformation model, or the warp model as it is traditionally coined in AAM, enables us to separate the shape and texture information of a given image by *warping* the texture to the mean shape. There a multiple ways of defining a warp using the points of the mean shape and the deformed shape as control points. However, it is standard practice to triangulate the mean shape and use linear interpolation to propagate vertex displacements to the points inside the simplexes. EAM adapts this approach.

The affine transformation between a simplex $\{\boldsymbol{v}_1, \dots \boldsymbol{v}_{N+1}\}$ and a simplex $\{\boldsymbol{v}_1 + \Delta\boldsymbol{v}_1, \dots, \boldsymbol{v}_{N+1} + \Delta\boldsymbol{v}_{N+1}\}$ can be defined by assuming that the point $\boldsymbol{x}$ inside $\{\boldsymbol{v}_1, \dots, \boldsymbol{v}_{N+1}\}$ corresponds to the point $\boldsymbol{\phi}(\boldsymbol{x})$ in $\{\boldsymbol{v}_1 + \Delta\boldsymbol{v}_1, \dots, \boldsymbol{v}_{N+1} + \Delta\boldsymbol{v}_{N+1}\}$ which has the same barycentric coordinates, see Figure H.2. The barycentric coordinates, $\boldsymbol{\eta}$, of an arbitrary point, $\boldsymbol{x}$, inside the simplex $\{\boldsymbol{v}_1, \dots \boldsymbol{v}_{N+1}\}$ are then given by

$$\begin{bmatrix} \boldsymbol{x} \\ 1 \end{bmatrix} = \begin{bmatrix} \boldsymbol{v}_1 & \cdots & \boldsymbol{v}_{N+1} \\ 1 & \cdots & 1 \end{bmatrix} \boldsymbol{\eta} = \begin{bmatrix} \boldsymbol{V} \\ \mathbf{1}^T \end{bmatrix} \boldsymbol{\eta}, \boldsymbol{\eta} = \begin{bmatrix} \boldsymbol{V} \\ \mathbf{1}^T \end{bmatrix}^{-1} \begin{bmatrix} \boldsymbol{x} \\ 1 \end{bmatrix} = \begin{bmatrix} \boldsymbol{D} & \boldsymbol{\eta}^0 \end{bmatrix} \begin{bmatrix} \boldsymbol{x} \\ 1 \end{bmatrix}, \tag{H.10}$$

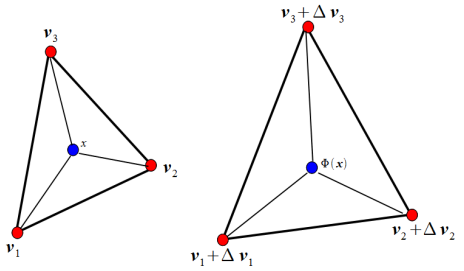


Figure H.2: Illustration of the affine transformation $\boldsymbol{\Phi}(\boldsymbol{x})$ from triangle (3-simplex) $\{\boldsymbol{v}_1, \boldsymbol{v}_2, \boldsymbol{v}_3\}$ to triangle $\{\boldsymbol{v}_1 + \Delta\boldsymbol{v}_1, \boldsymbol{v}_2 + \Delta\boldsymbol{v}_2, \boldsymbol{v}_3 + \Delta\boldsymbol{v}_3\}$.

where $\boldsymbol{\eta}^0$ is the barycentric coordinates for origo. Thus, the transformation between two simplices is given by

$$\boldsymbol{\phi}(\boldsymbol{x}) = \boldsymbol{x} + \Delta\boldsymbol{V}\boldsymbol{\eta} = \boldsymbol{x} + \Delta\boldsymbol{V}\begin{bmatrix}\boldsymbol{D} & \boldsymbol{\eta}^0\end{bmatrix}\begin{bmatrix}\boldsymbol{x}\\ 1\end{bmatrix}, \nabla\phi(\boldsymbol{x}) = \boldsymbol{I} + \Delta\boldsymbol{V}\boldsymbol{D}. \tag{H.11}$$

Note, as the transformation is linear the spatial derivatives of the transformation are independent of the spatial position within the fixed simplex. This means that the density energy in Equation H.8 is identical in all points within a simplex, and thus we only have to estimate Σ and $\bar{\boldsymbol{E}}_0$ once for every simplex instead of once for every pixel position in the shape-free reference space. The integrated elasticity energy for the entire deformation is

$$r(\Delta\boldsymbol{v}) = \sum_i^n a_i W_i(\log((\boldsymbol{I} + \Delta\boldsymbol{V}_i\boldsymbol{D}_i)^T(\boldsymbol{I} + \Delta\boldsymbol{V}_i\boldsymbol{D}_i))), \;\; a_i = \frac{1}{2}\left|\det\left(\begin{bmatrix}\boldsymbol{V}_i\\ \boldsymbol{1}^T\end{bmatrix}^{-1}\right)\right|, \tag{H.12}$$

where n is the number of simplices.

H.5.2 Active appearance model

In AAM, the model is matched to a template image, $\boldsymbol{T}$, by (see [35])

$$\underset{\boldsymbol{p}}{\operatorname{argmin}}\|\boldsymbol{g}(\boldsymbol{p}) - \boldsymbol{T}\circ\boldsymbol{\phi}(\boldsymbol{p})\|^2 + \alpha_1\|\boldsymbol{p}\|^2_{C^{-1}}, \tag{H.13}$$

where C is the covariance matrix of $\boldsymbol{p}$ and α_1 controls the strength of the parameter prior. α_1 is often set to zero and replaced by lower and upper bounds

on $\boldsymbol{p}$. As it, in practice, is impossible to initialize the matching perfectly with respect to pose, pose estimation is added to the optimization problem, i.e.

$$\underset{\boldsymbol{p},\boldsymbol{t}}{\operatorname{argmin}} \|\boldsymbol{g}(\boldsymbol{p}) - \boldsymbol{T} \circ \boldsymbol{\omega}(\boldsymbol{t}) \circ \boldsymbol{\phi}(\boldsymbol{p})\|^2 + \alpha_1 \|\boldsymbol{p}\|^2_{C^{-1}}, \tag{H.14}$$

where $\boldsymbol{\omega}$ is the similarity transformation.

H.5.3 Elastic appearance model

Similar to AAM, the EAM matches the appearance model to the template image using the sum of squared residuals with the addition of the statistical elasticity energy, i.e.

$$\underset{\boldsymbol{p},\Delta\boldsymbol{v}}{\operatorname{argmin}} \|\boldsymbol{g}(\boldsymbol{p}) - \boldsymbol{T} \circ \boldsymbol{\phi}(\Delta\boldsymbol{v})\|^2 + \alpha_1 \|\boldsymbol{p}\|^2_{C^{-1}} + \alpha_2 r(\Delta\boldsymbol{v}). \tag{H.15}$$

The advantages of Equation H.15 compared to Equation H.13-H.14 are that it cannot produce folding warps, and that it is not necessary to optimize rotation and translation separately due to local invariance. The disadvantage is that it is no longer a least square problem and thus requires the application of more advanced optimization algorithms. Our experience is that a limited-memory BFGS optimizer works well with the problem. Equation H.15 is not scale invariant which implies that Equation H.15 cannot match the shape model to an image if the scale between them is too large or small. One way to handle this is to introduce a global scaling parameter into the optimization. To avoid introducing a new parameter we prefer to filter out global scaling effects from the shape displacements prior to calling the regularizer, i.e.

$$\underset{\boldsymbol{p},\Delta\boldsymbol{v}}{\operatorname{argmin}} \|\boldsymbol{g}(\boldsymbol{p}) - \boldsymbol{T} \circ \boldsymbol{\phi}(\Delta\boldsymbol{v})\|^2 + \alpha_1 \|\boldsymbol{p}\|^2_{C^{-1}} + \alpha_2 r \left(\frac{\Omega_T(\bar{\boldsymbol{v}})}{\Omega_T(\bar{\boldsymbol{v}} + \Delta\boldsymbol{v})} (\bar{\boldsymbol{v}} + \Delta\boldsymbol{v}) - \bar{\boldsymbol{v}} \right), \tag{H.16}$$

where $\Omega_T(\boldsymbol{v})$ computes the area of a shape defined by the triangulation T and the vertices $\boldsymbol{v}$.

Even though Equation H.16 eliminates the need for separate pose estimation there are situations in practice where it is an advantage to restrict the displacement update to that of a similarity transformation – especially when the initial pose may be far from the true. On the other hand, if you are tracking an object frame-by-frame there is no need for such restrictions.

H.5.4 Implementation details

The AAM and EAM algorithms were implemented in Matlab using object oriented programming. Both the AAM and EAM use a limited-memory BFGS optimizer for the model matching. The AAM implementation also contains the traditional *linear relationship* optimization strategy proposed by Cootes *et al.* [33] but it was not used for the experiments in this paper as it in general performed inferior to the BFGS optimization. The minFunc and minConf optimization packages by Schmidt [135] is used in our implementation. Our implementation can be forwarded on request.

H.5.5 Experiments

The performance of the EAM was tested against the AAM in three experiments.

1) Face labeling: The capability of the AAM and EAM to model and label faces. This experiment used the subset of neutral facial expression images in the AR database [108] where annotation markup data were available from FGnet – in total 119 images. The data set images have a resolution of 768×576 with the mean distance of 111.2 pixels between eyes of the recorded persons. The performance of the methods was evaluated using leave-one-out cross-validation with the annotations as ground truth and the mean Euclidian point-to-point distance as the error measure. The AAM and EAM methods were initialized by aligning the mean shape to a set of eye coordinates found by a Viola-Jones face/eye detection [157]. As the Viola-Jones failed to detect the eyes in nine images, the experiment were carried out on 110 images. For future comparison of this work it should be noted that Viola-Jones detected eye positions poorly in three out of the 110 images. This was mainly do to heavy reflections in glasses on the three persons in question. This has no consequence on the performed experiments as both AAM and EAM are initially aligned with the same eye coordinates.

2) Face labeling with poor initialization: The robustness of the AAM and EAM against poor pose initialization. In this experiment the AAM and EAM methods were initialized several time by regular perturbation of the Viola-Jones alignment (experiment 1) in intervals $\sqrt{2}[-20pixel, 20pixel]$, $[-15°, 15°]$ and $[90\%, 110\%]$ for diagonal translation, rotation and scaling, respectively. 189 (9 translations, 7 rotations, 3 scales) registrations were performed per leave-one-out step. The performance of the methods was evaluated similarly to experiment 1.

3) Corpus Callosum segmentation: The performance of AAM and EAM on medical images. This experiment used 62 two dimensional MR images (with different subjects) of the mid-sagittal cross-section of the corpus callosum brain structure recorded at the Danish Reserach Centre for Magnetic Resonance, Hvidovre Hospital. This data set is part of the LADIS (Leukoaraiosis and DISability) study [118], a pan-European study involving 12 hospitals and more than 600 elderly subjects. Furthermore, each corpus callosum have manually been annotated with 72 landmarks by a clinician. The annotation were used as ground truth. The AAM and EAM methods were initialized several time by perturbation of the alignment between the mean shape and ground truth in the intervals $\sqrt{2}[-3pixel, 3pixel]$, $[-10°, 10°]$ and $[95\%, 105\%]$ for diagonal translation, rotation and scaling, respectively. The performance of the methods was evaluated with 6-fold cross-validation using the Dice, Jaccard and overlap coefficients to measure the quality of the segmentations. Each corpus callosum was segmented 75 times (5 translations, 5 rotations and 3 scales).

H.6 Experimental results

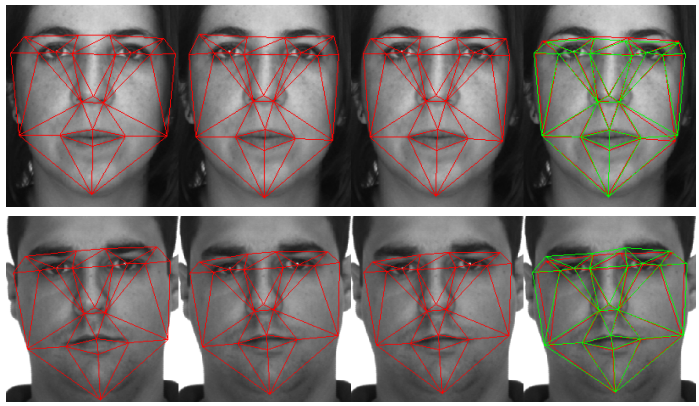
For all experiments principal component analysis was used for the matrix decomposition in Equation H.1-H.3 and the truncation level was 0.95 of the variance of the training set. Lower and upper bounds of 2.5 times the standard deviation were enforced on the model parameters and α_1 was set to zero. Practical experiments showed $\alpha_2 = 0.001$ for the faces and $\alpha_2 = 0.001$ for the corpus callosum gave good results.

H.6.1 Face labeling with Viola-Jones initialization

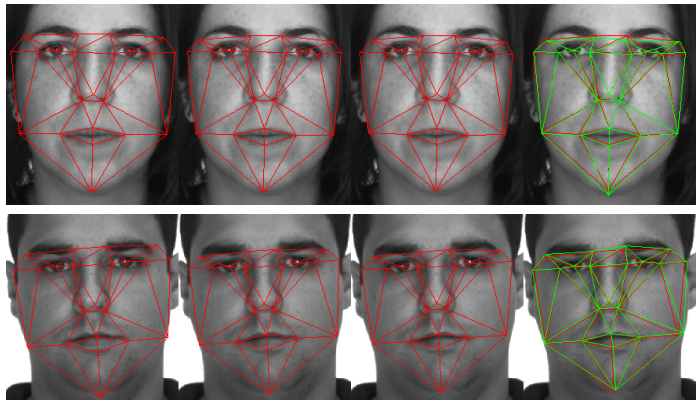
Table H.1 presents the results for the shape (annotation) displacement error. From the results it can be seen that EAM labels the face on average more than 10 per cent better than AAM. Visual examples of the face labeling are shown in Figure H.3, where an initial red triangulation of the shape is shown as it progresses to an optimum. The green triangulation shape shows the hand annotated ground truth.

Method	Displacement error in pixel			
	Mean	25th percentile	75th percentile	97.5th percentile
AAM	4.8	3.1	4.7	15.2
EAM	4.3	3.1	4.5	8.8

Table H.1: Displacement error, initialized with Viola-Jones face/eye detection.



(a) AAM



(b) EAM

Figure H.3: Progress left-to-right from initialized mean shape with Viola-Jones face/eye detection to converged result for AAM and EAM, respectively. The red triangulation shows the shape as it progresses to an optimum. The green triangulation signifies the hand annotated ground truth.

H.6.2 Face labeling with poor initialization

When making a poor initialization of the models it can be seen that EAM really outperforms AAM. By labeling the face on average 39 per cent better than AAM. Table H.2 presents the displacement error results. Visual examples of the face labeling are shown in Figure H.4.

Method	Displacement error in pixel			
	Mean	25th percentile	75th percentile	97.5th percentile
AAM	13.1	3.9	16.8	52.8
EAM	8.0	3.3	7.8	34.0

Table H.2: Displacement error, initialized with poor face detection.

H.6.3 Corpus Callosum segmentation

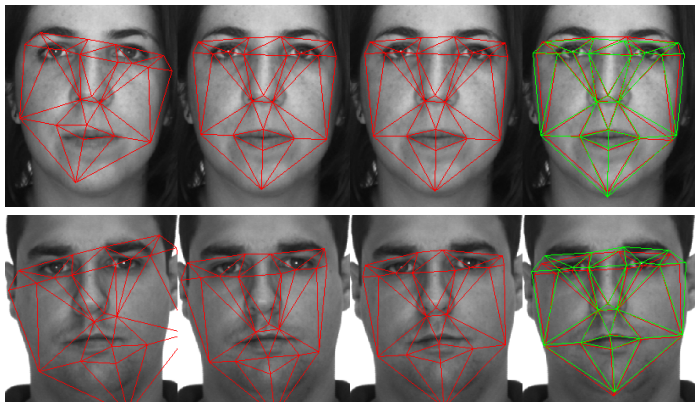
The results of the experiment are listed in Table H.3 and visualized in Figure H.5. As the Dice and Jaccard coefficients after matching are 0.89 and 0.81 for AAM and 0.90 and 0.82 for EAM using only the unperturbed ground truth alignments for initialization the large differences in Dice and Jaccard coefficients primarily indicate that the EAM is more robust and secondarily that is it more accurate. This conclusion is supported by the histograms of the coefficients shown in Figure H.5.

Method	Mean		
	Dice	Jaccard	Overlap
AAM	0.75	0.63	0.85
EAM	0.80	0.70	0.86

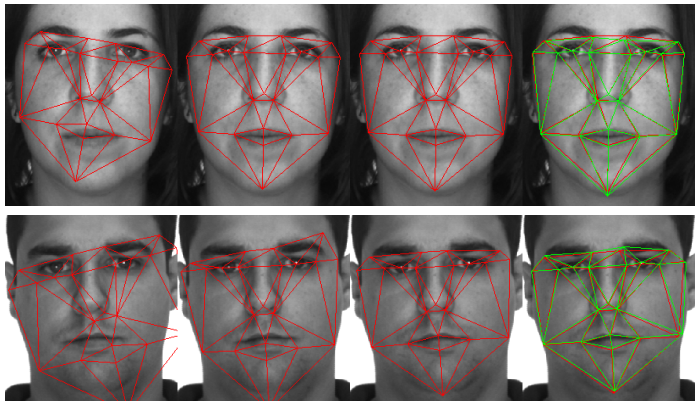
Table H.3: Means for the Dice, Jaccard and overlap coefficients for Corpus Callosum segmentation.

H.7 Discussion and future work

The experimental results favors EAM against AAM. However, as the implementations of AAM and EAM are basic in the sense, that they do not incorporate any of the improvement mentioned in the previous work section, the difference in performance may be smaller as soon as the implementations are combined

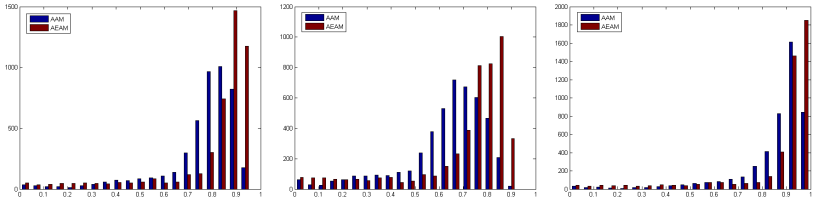


(a) AAM



(b) EAM

Figure H.4: Progress left-to-right from initialized mean shape with poor face detection to converged result for AAM and EAM, respectively. The red triangulation shows the shape as it progresses to an optimum. The green triangulation signifies the hand annotated ground truth.



(a) Dice (b) Jaccard (c) Overlap

Figure H.5: Histograms of the Dice, Jaccard and overlap coefficients for Corpus Callosum.

with some of the other improvements. Still, we believe that the experiments are good performance indicators.

The purpose of many of the AAM improvements is to speed-up the matching with real-time tracking being the end goal. It can be argued that we have gained robustness and accuracy with the EAM but lost efficiency as the problem in Equation H.16 is much less cost efficient than e.g. the linear relationship strategy proposed by Cootes *et al.* [33]. However, if the application is real-time frame-by-frame tracking of an object the EAM formulation may not be as computational expensive as it seems. First of all it is not necessary to do pose estimation as long as the object movement is slow compared to the framerate. Secondly, if you apply a BFGS optimizer you can use the estimate of the Hessian of the previous frame as initial estimate of cost function hessian which means that you may only need a small number of iterations to get to the optimum.

To train the shape model in EAM you need to estimate a small 3x3 strain covariance matrix for each triangle in the model (Equation H.9) while you have to estimate a large covariance matrix and decompose it in PCA based AAM. A consequence of this is that fewer samples are needed to train the EAM than the AAM.

Bibliography

[1] *International Statistical Classification of Diseases and Related Health Problems - ICD 10.* World Health Organization, 10th edition, 2010.

[2] B. Abboud, F. Davoine, and M. Dang. Facial expression recognition and synthesis based on an appearance model. *Signal Processing: Image Communication*, 19(8):723–740, 2004.

[3] M. Aizerman, E. Braverman, and L. Rozonoer. Theoretical foundations of the potential function method in pattern recognition learning. *Automation and Remote Control*, 25:821–837, 1964.

[4] J.D. Andersen, P. Johansen, S. Harder, S. Christoffersen, M. Delgado, S. Henriksen, M. Nielsen, E. Sørensen, et al. Genetic analyses of the human eye colour using a novel objective method for eye colour classification. *Forensic Science International: Genetics*, 7:508–515, 2013.

[5] E. Bair, T. Hastie, D. Paul, and R. Tibshirani. Prediction by supervised principal components. *Journal of the American Statistical Association*, 101:119–137, 2006.

[6] S. Baker and I. Matthews. Lucas-kanade 20 years on: A unifying framework. *International Journal of Computer Vision*, 56(3):221–255, 2004.

[7] S.E. Baranzini, J. Mudge, J.C. van Velkinburgh, P. Khankhanian, I. Khrebtukova, N.A. Miller, L. Zhang, A.D. Farmer, et al. Genome, epigenome and RNA sequences of monozygotic twins discordant for multiple sclerosis. *Nature*, 464(7293):1351–1356, 2010.

[8] J. Barbeito-Andrés, M. Anzelmo, F. Ventrice, and M.L. Sardi. Measurement error of 3D cranial landmarks of an ontogenetic sample using com-

puted tomography. *Journal of Oral Biology and Craniofacial Research*, 2(2):77–82, 2012.

[9] D. Baucum. *Psychology.* Barron's, 2nd edition, 2006.

[10] R. Beichel, G. Gotschuli, E. Sorantin, F.W. Leberl, and M. Sonka. Diaphragm dome surface segmentation in CT data sets: A 3D active appearance model approach. *Proceedings of SPIE Medical Imaging 2002*, 4684:475, 2002.

[11] J. Bekios-Calfa, J.M. Buenaposada, and L. Baumela. Revisiting linear discriminant techniques in gender recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(4):858–864, 2011.

[12] P.N. Belhumeur, J.P. Hespanha, and D.J. Kriegman. Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19:711–720, 1997.

[13] R. Bellman. *Adaptive Control Processes: A Guided Tour.* Princeton University Press, 1961.

[14] K. Beyer, J. Goldstein, R. Ramakrishnan, and U. Shaft. When is "nearest neighbor" meaningful? In *In Int. Conf. on Database Theory*, pages 217–235, 1999.

[15] I. Biederman and P. Kalocsai. Neurocomputational bases of object and face recognition. *Philosophical Transactions: Biological Sciences*, 352(1358):1203–1219, 1997.

[16] V. Blanz and T. Vetter. Face recognition based on fitting a 3D morphable model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(9):1063–1074, 2003.

[17] S. Boehringer, F. van der Lijn, F. Liu, M. Gunther, S. Sinigerova, S. Nowak, K.U. Ludwig, R. Herberz, et al. Genetic determination of human facial morphology: links between cleft-lips and normal variation. *European Journal of Human Genetics*, 19(11):1192–1197, 2011.

[18] F.L. Bookstein. Principal warps: Thin-plate splines and the decomposition of deformations. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 11(6):567–585, 1989.

[19] D.H. Brainard. The psychophysics toolbox. *Spatial Vision*, 10(4):433–436, 1997.

[20] V. Bruce, A. Burton, E. Hanna, P. Healey, O. Mason, A. Coombes, R. Fright, and A. Linney. Sex discrimination: how do we tell the difference between male and female faces? *Perception*, 22(2):131–152, 1993.

[21] N. Brunetti-Pierri, J.S. Berg, F. Scaglia, J. Belmont, C.A. Bacino, T. Sahoo, S.R. Lalani, B. Graham, et al. Recurrent reciprocal 1q21.1 deletions and duplications associated with microcephaly or macrocephaly and developmental and behavioral abnormalities. *Nature Genetics*, 40(12):1466–1471, Dec 2008.

[22] C.J.C. Burges. A Tutorial on Support Vector Machines for Pattern Recognition. *Data Mining and Knowledge Discovery*, pages 121–167, 1998.

[23] C.J.C. Burges. Dimension Reduction: A Guided Tour. *Foundations and Trends in Machine Learning*, 2(4):275–364, 2009.

[24] P.H. Buschang, R. Tanguay, and A. Demirjian. Cephalometric reliability: a full ANOVA model for the estimation of true and error variance. *The Angle Orthodontist*, 57(2):168–175, 1987.

[25] W.S. Bush and J.H. Moore. Chapter 11: Genome-wide association studies. *PLoS computational biology*, 8(12):e1002822, 2012.

[26] F. Camastra. Data dimensionality estimation methods: A survey. *Pattern Recognition*, 36:2945–2954, 2003.

[27] H. Cevikalp, M. Neamtu, M. Wilkes, and A. Barkana. Discriminative common vectors for face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27:4–13, 2005.

[28] V. Cherkassky and Mulier F.M. *Learning from Data: Concepts, Theory, and Methods*. John Wiley and Sons, 1998.

[29] K. Chinthapalli, E. Bartolini, J. Novy, M. Suttie, C. Marini, M. Falchi, Z. Fox, L. M S Clayton, et al. Atypical face shape and genomic structural variants in epilepsy. *Brain : a journal of neurology*, 135(10):3101–3114, 2012.

[30] C.M. Christoudias and T. Darrell. On modelling nonlinear shape-and-texture appearance manifolds. *Proceedings of the 2005 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1067–1074, 2005.

[31] E.H. Cook and S.W. Scherer. Copy-number variations associated with neuropsychiatric conditions. *Nature*, 455(7215):919–923, Oct 2008.

[32] G.M. Cooper, B.P. Coe, S. Girirajan, J.A. Rosenfeld, T.H. Vu, C. Baker, C. Williams, H. Stalker, et al. A copy number variation morbidity map of developmental delay. *Nature Genetics*, 43(9):838–846, Sep 2011.

[33] T.F. Cootes, G.J. Edwards, and C.J. Taylor. Active Appearance Models. *Proceedings of the 5th European Conference on Computer Vision (LNCS)*, 1407:484–498, 1998.

[34] T.F. Cootes, G.J. Edwards, and C.J. Taylor. Active Appearance Models. *IEEE Transactions on. Pattern Analysis and Machine Intelligence*, 23(6):681–685, 2001.

[35] T.F. Cootes and C.J. Taylor. Constrained active appearance models. *Proceedings of the 8th International Conference on Computer Vision*, 2001.

[36] T.F. Cootes and C.J. Taylor. Statistical models of appearance for computer vision. Technical report, Imaging Science and Biomedical Engineering. University of Manchester, March 2004.

[37] T.F. Cootes, C.J. Taylor, D.H. Cooper, and J. Graham. Active Shape Models-Their Training and Application. *Computer Vision and Image Understanding*, 61(1):38–59, 1995.

[38] H.S.M. Coxeter. *Introduction to geometry (2nd ed.)*. John Wiley and Sons, 1969.

[39] J. Dahl. Analysis pipeline for face recognition. *B.Sc.Eng. Thesis*, Technical University of Denmark, 2012.

[40] R.H. Davies, C.J. Twining, T.F. Cootes, J.C. Waterton, and C.J. Taylor. A Minimum Description Length Approach to Statistical Shape Modeling. *IEEE Transactions on. Medical Imaging*, 21(5):525–537, 2002.

[41] R.H. Davies, C.J. Twining, P. Daniel Allen, T.F. Cootes, and C.J. Taylor. Building optimal 2D statistical shape models. *Image and Vision Computing*, 21:1171–1182, 2003.

[42] N. de Souza. SNPing away at anonymity. *Nature Methods*, pages 918–919, 2008.

[43] D.L. Donoho. Aide-memoire. high-dimensional data analysis: The curses and blessings of dimensionality, 2000.

[44] M. Doornbos, B. Sikkema-Raddatz, C.A. Ruijvenkamp, T. Dijkhuizen, E.K. Bijlsma, A.C. Gijsbers, Y. Hilhorst-Hofstee, R. Hordijk, et al. Nine patients with a microdeletion 15q11.2 between breakpoints 1 and 2 of the Prader-Willi critical region, possibly associated with behavioural disturbances. *European Journal of Medical Genetics*, 52(2-3):108–115, 2009.

[45] L. Duong. Mutations in NRXN1-Locus in a family with diverse psychopathological phenotypes. *Schizophrenia research submitted*, 2012.

[46] L. Duong, L.L. Klitten, R.S. Møller, A. Ingason, K.D. Jakobsen, C. Skjodt, M. Didriksen, H. Hjalgrim, T. Werge, and N. Tommerup. Mutations in NRXN1 in a family multiply affected with brain disorders: NRXN1 mutations and brain disorders. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics*, 159B(3):354–358, Apr 2012.

[47] G.J. Edwards, T.F. Cootes, and C.J. Taylor. Face recognition using active appearance models. *Proceedings of the 5th European Conference on Computer Vision (LNCS)*, page 581, 1998.

[48] H. Eiberg, J. Troelsen, M. Nielsen, A. Mikkelsen, J. Mengel-From, K.W. Kjaer, and L. Hansen. Blue eye color in humans may be caused by a perfectly associated founder mutation in a regulatory element located within the HERC2 gene inhibiting OCA2 expression. *Human Genetics*, 123(2):177–187, 2008.

[49] N. Eriksson, J.M. Macpherson, J.Y. Tung, L.S. Hon, B. Naughton, S. Saxonov, L. Avey, A. Wojcicki, I. Pe'er, and J. Mountain. Web-based, participant-driven studies yield novel genetic associations for common traits. *PLoS Genetics*, 6(6):e1000993, 2010.

[50] J. Fagertun. *Face Recognition. M.Sc.Eng. Thesis*, Technical University of Denmark, 2005.

[51] J. Fagertun, T. Andersen, T.F. Hansen, and R.R. Paulsen. 3D gender recognition using cognitive modeling. In *2013 International Workshop on Biometrics and Forensics (IWBF)*. IEEE, 2013.

[52] J. Fagertun, T. Andersen, and R.R. Paulsen. Gender recognition using cognitive modeling. In A. Fusiello, V. Murino, and R. Cucchiara, editors, *Computer Vision – ECCV 2012. Workshops and Demonstrations*, volume 7584 of *Lecture Notes in Computer Science*, pages 300–308. Springer Berlin Heidelberg, 2012.

[53] J. Fagertun, D.D. Gomez, B.K. Ersbøll, and R. Larsen. *A face recognition algorithm based on multiple individual discriminative models*, pages 69–75. DIKU, University of Copenhagen, 2005.

[54] J. Fagertun, D.D. Gomez, M.F. Hansen, and R.R. Paulsen. Sparse similarity-based fisherfaces. In A. Heyden and F. Kahl, editors, *Image Analysis*, volume 6688 of *Lecture Notes in Computer Science*, pages 69–78. Springer Berlin Heidelberg, 2011.

[55] J. Fagertun, S. Harder, A. Rosengren, C.K. Møller, T. Werge, R.R. Paulsen, and T.F. Hansen. Variability of 3D facial landmarks. *submitted to Medical Image Analysis*, 2013.

[56] J. Fagertun, C.K. Møller, T. Werge, R.R. Paulsen, and T.F. Hansen. Detection of minor physical anomalies in the face associated with severe schizophrenic disorders. In *submission*.

[57] J. Fagertun, T. Pers, K. Wolffhechel, R.R. Paulsen, and H. Jarmer. Correlation of complex polygenic variations to specific holistic facial characteristics. *submitted to PLoS ONE*, 2014.

[58] J. Fagertun and M.B. Stegmann. The IMM frontal face database. Technical report, Informatics and Mathematical Modelling, Technical University of Denmark, DTU, 2005.

[59] H. Faulds. On the skin-furrows of the hand. *Nature*, 22:605, October 1880.

[60] L. Flintoft. Enhancers 'fine-tune' face and skull shape. *Nature Reviews Genetics*, 14, 2013.

[61] M. Frisenfeldt. 3D reconstruction of the human face. *B.Sc.Eng. Thesis*, Technical University of Denmark, 2013.

[62] F. Galton. *Finger Prints.* Macmillan London, 1892.

[63] X. Gao, Y. Su, X. Li, and D. Tao. A review of active appearance models. *IEEE Transactions on Systems, Man, and Cybernetics*, 40(2):145–158, 2010.

[64] I. Gauthier and M. Behrmann. Can face recognition really be dissociated from object recognition? *Journal of Cognitive Neuroscience*, 11(4), 1999.

[65] A. Ghodsi. Dimensionality Reduction A Short Tutorial. Technical report, Department of Statistics and Actuarial Science, Univ. of Waterloo, Ontario, 2006.

[66] S. Girirajan, Z. Brkanac, B.P. Coe, C. Baker, L. Vives, T.H. Vu, N. Shafer, R. Bernier, et al. Relative burden of large CNVs on a range of neurodevelopmental phenotypes. *PLoS Genetics*, 7(11):e1002334, Nov 2011.

[67] J.T. Glessner, K. Wang, G. Cai, O. Korvatska, C.E. Kim, H. Zhang, A. Estes, C.W. Brune, et al. Autism genome-wide copy number variation reveals ubiquitin and neuronal genes. *Nature*, 459(7246):569–573, 2010.

[68] D.D. Gomez, J. Fagertun, B. Ersbøll, F.M. Sukno, and A.F. Frangi. Similarity-based fisherfaces. *Pattern Recognition Letters*, 30(12):1110–1116, 2009.

[69] C. Goodall. Procrustes methods in the statistical analysis of shape. *Journal of the Royal Statistical Society*, 53(2):285–339, 1991.

[70] H. Gould and J. Tobochnik. *Statistical and Thermal Physics: With Computer Applications.* Princeton University Press, 2010.

[71] R. Gross, I. Matthews, and S. Baker. Appearance-based face recognition and light-fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(4):449–465, 2004.

[72] J. R. Gulcher, K. Kristjansson, H. Gudbjartsson, and K. Stefansson. Protection of privacy by third-party encryption in genetic research in Iceland. *European Journal of Human Genetics*, 8(10):739–742, 2000.

[73] L. Guyot, M. Dubuc, J. Pujol, O. Dutour, and N. Philip. Craniofacial anthropometric analysis in patients with 22q11 microdeletion. *American Journal of Medical Genetics*, 100(1):1–8, Apr 2001.

[74] P. Hammond, T.J. Hutton, J.E. Allanson, L.E. Campbell, R.C.M. Hennekam, S. Holden, M.A. Patton, A. Shaw, et al. 3D analysis of facial morphology. *American journal of medical genetics. Part A*, 126A(4):339–348, 2004.

[75] D.W. Hansen, J.P. Hansen, M. Nielsen, A.S. Johansen, and M.B. Stegmann. Eye typing using Markov and active appearance models. *Proceedings of the sixth IEEE Workshop on Applications of Computer Vision*, pages 132–136, 2002.

[76] M.F. Hansen, J. Fagertun, and L. Larsen. Elastic appearance models. In *Proceedings of the British Machine Vision Conference*. BMVA Press, 2011.

[77] T.F. Hansen, J. Fagertun, L. Duong, A. Rosengren, E. Sørensen, H. Ullum, R. Larsen, R.R. Paulsen, and T. Werge. Discrete facial feature of NRNX1 deletions carriers. In *submission*.

[78] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2009.

[79] X. He, S. Yan, Y. Hu, P. Niyogi, and H. Zhang. Face recognition using Laplacianfaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(3):328–340, 2005.

[80] J. Heikkila and O. Silven. A four-step camera calibration procedure with implicit image correction. *Computer Vision and Pattern Recognition*, pages 1106–1112, 1997.

[81] G. Hemani, J. Yang, A. Vinkhuyzen, J.E. Powell, G. Willemsen, J.J. Hottenga, A. Abdellaoui, M. Mangino, et al. Inference of the Genetic Architecture Underlying BMI and Height with the Use of 20,240 Sibling Pairs. *American Journal of Human Genetics*, 93(5):865–875, 2013.

[82] K.B. Hilger, R.R. Paulsen, and R. Larsen. Markov random field restoration of point correspondences for active shape modeling. In *Medical Imaging 2004*, pages 1862–1869. International Society for Optics and Photonics, 2004.

[83] I.I. Hirschman and D.V. Widder. *The Convolution Transform*. Princeton University Press, 1955.

[84] H. Holm, D. F. Gudbjartsson, P. Sulem, G. Masson, H. T. Helgadottir, C. Zanon, O. T. Magnusson, A. Helgason, et al. A rare variant in MYH6 is associated with high risk of sick sinus syndrome. *Nature Genetics*, 43(4):316–320, 2011.

[85] G. Hughes. On the mean accuracy of statistical pattern recognizers. *IEEE Transactions on Information Theory*, 14(1):55–63, 1968.

[86] T.J. Hutton, B.R. Buxton, and P. Hammond. Dense surface point distribution models of the human face. *Proceedings IEEE Workshop on Mathematical Methods in Biomedical Image Analysis (MMBIA 2001)*, pages 153–160, 2001.

[87] A.J. Iafrate, L. Feuk, M.N. Rivera, M.L. Listewnik, P.K. Donahoe, Y. Qi, S.W. Scherer, and C. Lee. Detection of large-scale variation in the human genome. *Nat. Genet.*, 36(9):949–951, 2004.

[88] A.K. Jain, P. Flynn, and A.A. Ross. *Handbook of Biometrics*. Springer, 2008.

[89] A.K. Jain, A. Ross, and S. Prabhakar. An introduction to biometric recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 14(1):4–20, 2004.

[90] M. Journee, Y. Nesterov, P. Richtarik, and R. Sepulchre. Generalized power method for sparse principal component analysis. *The Journal of Machine Learning Research*, 11:517–553, 2010.

[91] F. Kahraman, M.B. Stegmann, S. Darkner, and R. Larsen. An Active Illumination and Appearance (AIA) Model for Face Alignment. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,2007*, pages 1–7, 2007.

[92] M. Kayser, F. Liu, A. C. Janssens, F. Rivadeneira, O. Lao, K. van Duijn, M. Vermeulen, P. Arp, et al. Three genome-wide association studies and a linkage analysis identify HERC2 as a human iris color gene. *American Journal of Human Genetics*, 82(2):411–423, 2008.

[93] B. Keating, A.T. Bansal, S. Walsh, J. Millman, J. Newman, K. Kidd, B. Budowle, A. Eisenberg, et al. First all-in-one diagnostic tool for DNA intelligence: genome-wide inference of biogeographic ancestry, appearance, relatedness, and sex with the Identitas v1 Forensic Chip. *International Journal of Legal Medicine*, 127(3):559–572, 2013.

[94] E.E. Kenny, N.J. Timpson, M. Sikora, M.C. Yee, A. Moreno-Estrada, C. Eng, S. Huntsman, E.G. Burchard, et al. Melanesian blond hair is caused by an amino acid change in TYRP1. *Science*, 336(6081):554, 2012.

[95] M. Kirby and L. Sirovich. Application of the karhunen-Loeve procedure for the characterization of human faces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12:103–108, 1990.

[96] H. Lango Allen, K. Estrada, G. Lettre, S.I. Berndt, M.N. Weedon, F. Rivadeneira, C.J. Willer, A.U. Jackson, et al. Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature*, 467(7317):832–838, 2010.

[97] R. Larsen and K. Baggesen. Statistical shape analysis using Non-Euclidean Metrics. *Medical Image Analysis*, 7(4):417–423, 2003.

[98] R. Larsen, M.B. Stegmann, S. Darkner, S. Forchhammer, T.F. Cootes, and B.K. Ersboll. Texture enhanced appearance models. *Computer Vision and Image Understanding*, 106(1):20–30, 2007.

[99] D. Leimberg and M. Vester-Christensen. Eye tracking. *M.Sc.Eng. Thesis*, Technical University of Denmark, 2005.

[100] S.Z. Li. *Markov random field modeling in image analysis*. Springer, 2009.

[101] P. Lichtenstein, B.H. Yip, C. Bjork, Y. Pawitan, T.D. Cannon, P.F. Sullivan, and C.M. Hultman. Common genetic determinants of schizophrenia and bipolar disorder in Swedish families: a population-based study. *Lancet*, 373(9659):234–239, 2009.

[102] F. Liu, F. van der Lijn, C. Schurmann, G. Zhu, M.M. Chakravarty, P.G. Hysi, A. Wollstein, O. Lao, et al. A genome-wide association study identifies five loci influencing facial morphology in Europeans. *PLoS Genetics*, 8(9), 2012.

[103] Q. Liu, R. Huang, H. Lu, and S. Ma. Face recognition using kernel-based fisher discriminant analysis. *Proceedings: Fifth IEEE International Conference on Automatic Face and Gesture Recognition*, pages 197–201, 2002.

[104] S. Maiti, K.H. Kumar, C.A. Castellani, R. O'Reilly, and S.M. Singh. Ontogenetic de novo copy number variations (CNVs) as a source of genetic individuality: studies on two families with MZD twins for schizophrenia. *PLoS ONE*, 6(3):e17125, 2011.

[105] W. Majewski. Identification of the optimal method for analysis of human faces in the context of finding associations to personality traits. *Special Course*, Technical University of Denmark, 2012.

[106] E. Mäkinen and R. Raisamo. Evaluation of gender classiffcation methods with automatically detected and aligned faces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(3):541–547, 2008.

[107] T.A. Manolio, F.S. Collins, N.J. Cox, D.B. Goldstein, L.A. Hindorff, D.J. Hunter, M.I. McCarthy, E.M. Ramos, et al. Finding the missing heritability of complex diseases. *Nature*, 461(7265):747–753, 2009.

[108] A.M. Martinez and R. Benavente. The AR face database. *CVC Technical Report* 24, Computer Vision Center Purdue University, 1998.

[109] M. McGue and I.I. Gottesman. The genetic epidemiology of schizophrenia and the design of linkage studies. *European Archives of Psychiatry and Clinical Neuroscience*, 240(3):174–181, 1991.

[110] S.E. Medland, D.R. Nyholt, J.N. Painter, B.P. McEvoy, A.F. McRae, G. Zhu, S.D. Gordon, M.A. Ferreira, et al. Common variants in the trichohyalin gene are associated with straight hair in Europeans. *American Journal of Human Genetics*, 85(5):750–755, 2009.

[111] H.C. Mefford, A.J. Sharp, C. Baker, A. Itsara, Z. Jiang, K. Buysse, S. Huang, V.K. Maloney, et al. Recurrent rearrangements of chromosome 1q21.1 and variable pediatric phenotypes. *New England Journal of Medicine*, 359(16):1685–1699, 2008.

[112] K. Messer, J. Matas, J. Kittler, J. Luettin, and G. Maitre. XM2VTSDB: The extended M2VTS database. *Second International Conference on Audio and Videobased Biometric Person Authentication*, 1999.

[113] S.C. Mitchell, B.P.F. Lelieveldt, R.J. van der Geest, H.G. Bosch, J.H.C. Reiver, and M. Sonka. Multistage hybrid active appearance model matching: segmentation of left and right ventricles in cardiac MR images. *IEEE Transactions on Medical Imaging*, 20(5):415–423, 2001.

[114] B. Moghaddam and M.-H. Yang. Learning gender with support faces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(5):707–711, 2002.

[115] C.K. Møller. 3-Dimentionel craniofacial image analysis: Detection of anthropometric traits as a marker for genetically caused disease risk. *B.Sc.Eng. Thesis*, Copenhagen University Hospital, 2013.

[116] C.M. Mooney. Age in the development of closure ability in children. *Canadian Journal of Psychology*, 11(4):219–226, 1957.

[117] S. Oskarsdottir, C. Persson, B.O. Eriksson, and A. Fasth. Presenting phenotype in 100 children with the 22q11 deletion syndrome. *European Journal of Pediatrics*, 164(3):146–153, Mar 2005.

[118] L. Pantoni, A.M. Basile, G. Pracucci, K. Asplund, J. Bogousslavsky, H. Chabriat, T. Erkinjuntti, F. Fazekas, et al. Impact of age-related cerebral white matter changes on the transition to disability–the LADIS study: rationale, design and methodology. *Neuroepidemiology*, 24(1-2):51–62, 2004.

[119] L. Paternoster, A.I. Zhurov, A.M. Toma, J.P. Kemp, B. St Pourcain, N.J. Timpson, G. McMahon, W. McArdle, et al. Genome-wide association study of three-dimensional facial morphology identifies a variant in PAX3 associated with nasion position. *American Journal of Human Genetics*, 90(3):478–485, 2012.

[120] R.R. Paulsen. *Statistical shape analysis of the human ear canal with application to in-the-ear hearing aid design.* IMM, Informatik og Matematisk Modellering, Danmarks Tekniske Universitet, 2004.

[121] R.R. Paulsen and K.B. Hilger. Shape modelling using markov random field restoration of point correspondences. In *Information Processing in Medical Imaging*, pages 1–12. Springer, 2003.

[122] O.B. Pedersen, C. Erikstrup, S.R. Kotzé, E. Sørensen, M.S. Petersen, K. Grau, and H. Ullum. The Danish Blood Donor Study: a large, prospective cohort and biobank for medical research. *Vox Sanguinis*, 102(3):271, 2012.

[123] X. Pennec, R. Stefanescu, V. Arsigny, P. Fillard, and N. Ayache. Riemannian Elasticity: A Statistical Regularization Framework for Nonlinear Registration. *Proceedings of the 8th International Conference on Medical Image Computing and Computer-Assisted Interventation (LNCS)*, 3750:943–950, 2005.

[124] H. Phillips, P. Moon, and S. Rizvi. The FERET evaluation methodology for face recognition algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(10), 2000.

[125] S.M. Pizer, P.T. Fletcher, Y. Fridman, D.D. Fritsch, A.G. Gash, J.M. Glotzer, S. Joshi, A. Thall, et al. Deformable m-reps for 3d medical image segmentation. *International Journal of Computer Vision*, 55(2):85–106, 2003.

[126] J.R. Pollack, C.M. Perou, A.A. Alizadeh, M.B. Eisen, A. Pergamenschikov, C.F. Williams, S.S. Jeffrey, D. Botstein, and P.O. Brown. Genome-wide analysis of DNA copy-number changes using cDNA microarrays. *Nat. Genet.*, 23(1):41–46, 1999.

[127] F.P. Preparata and M.I. Shamos. *Computational Geometry: An Introduction.* Springer, 1985.

[128] D.E. Reich and E.S. Lander. On the allelic spectrum of human disease. *Trends in genetics : TIG*, 17(9):502–10, 2001.

[129] L.E. Rosenberg and D.D. Rosenberg. *Human Genes and Genomes: Science, Health, Society.* Academic Press, first edition, 2012.

[130] A. Ross and A.K. Jain. Multimodal biometrics: An overview. *Proc. of 12th European Signal Processing Conference (EUSIPCO)*, pages 1221–1224, 2004.

[131] D. Rujescu, A. Ingason, S. Cichon, O.P.H. Pietiläinen, R. Michael, T. Toulopoulou, M. Picchioni, E. Vassos, et al. Disruption of the neurexin 1 gene is associated with schizophrenia. *Human Molecular Genetics*, 18(5):988–996, 2009.

[132] S. Sahin and J.H. Saabye. The connection between 3D facial scans and schizophrenia. *Special Course*, Technical University of Denmark, 2011.

[133] P.S. Sanjekar and J.B. Patil. An overview of multimodal biometrics. *Signal and Image Processing : An International Journal*, 4(1), 2013.

[134] J. Schiffman, M. Ekstrom, J. LaBrie, F. Schulsinger, H. Sorensen, and S. Mednick. Minor physical anomalies and schizophrenia spectrum disorders: a prospective investigation. *American Journal of Psychiatry*, 159(2):238–243, Feb 2002.

[135] M. Schmidt. minfunc and minconf optimization package, 2009. `http://www.cs.ubc.ca/~schmidtm/Software/minConf.html`.

[136] I.M. Scott, T. Cootes, and C. Taylor. Improving appearance model matching using local image structure. *Proceedings of Information Processing in Medical Imaging, 2003*, pages 258–269, 2003.

[137] J. Sebat, B. Lakshmi, J. Troge, J. Alexander, J. Young, P. Lundin, S. Maner, H. Massa, et al. Large-scale copy number polymorphism in the human genome. *Science*, 305(5683):525–528, 2004.

[138] A. Søgaard and S. Bohøj. 3D feature extraction of 3D human facials scans. *B.Sc.Eng. Thesis*, Technical University of Denmark, 2012.

[139] A.J. Sharp, H.C. Mefford, K. Li, C. Baker, C. Skinner, R.E. Stevenson, R.J. Schroer, F. Novara, et al. A recurrent 15q13.3 microdeletion syndrome associated with mental retardation and seizures. *Nature Genetics*, 40(3):322–328, Mar 2008.

[140] S.B. Sholts, L. Flores, P.L. Walker, and S.K.T.S. Wärmländer. Comparison of coordinate measurement precision of different landmark types

on human crania using a 3D laser scanner and a 3D digitiser: implications for applications of digital morphometrics. *International Journal of Osteoarchaeology*, 21(5):535–543, 2011.

[141] S.M. Silverstein, W.D. Spaulding, and A.A. Menditto. *Schizophrenia - Advances in Psychotherapy*. Hogrefe, 1st edition, 2006.

[142] H. Stefansson, A. Meyer-Lindenberg, S. Steinberg, B. Magnusdottir, K. Morgen, S. Arnarsdottir, G. Bjornsdottir, G. B. Walters, et al. CNVs conferring risk of autism or schizophrenia affect cognition in controls. *Nature*, 505(7483):361–366, 2014.

[143] H. Stefansson, D. Rujescu, S. Cichon, O.P.H. Pietiläinen, A. Ingason, S. Steinberg, R. Fossdal, E. Sigurdsson, et al. Large recurrent microdeletions associated with schizophrenia. *Nature*, 455(7210):232–6, 2008.

[144] M.B. Stegmann. Object tracking using active appearance models. *Proceedings of the 10th Danish Conference on Pattern Recognition and Image Analysis*, 1:54–60, 2001.

[145] M. Steinbach, L. Ertöz, and V. Kumar. The challenges of clustering high-dimensional data. In *In New Vistas in Statistical Physics: Applications in Econophysics, Bioinformatics, and Pattern Recognition*. Springer-Verlag, 2003.

[146] J.L. Stone, M.C. O'Donovan, H. Gurling, G.K. Kirov, D.H. Blackwood, A. Corvin, N.J. Craddock, M. Gill, et al. Rare chromosomal deletions and duplications increase risk of schizophrenia. *Nature*, 455(7210):237–241, Sep 2008.

[147] R.A. Sturm, D.L. Duffy, Z.Z. Zhao, F.P. Leite, M.S. Stark, N.K. Hayward, N.G. Martin, and G.W. Montgomery. A single SNP in an evolutionary conserved region within intron 86 of the HERC2 gene determines human blue-brown eye color. *American Journal of Human Genetics*, 82(2):424–431, Feb 2008.

[148] N. Sugano. Computer-assisted orthopedic surgery. *Journal of Orthopaedic Science*, 8(3):442–448, 2003.

[149] P. Sulem, D.F. Gudbjartsson, S.N. Stacey, A. Helgason, T. Rafnar, K.P. Magnusson, A. Manolescu, A. Karason, et al. Genetic determinants of hair, eye and skin pigmentation in Europeans. *Nature Genetics*, 39(12):1443–1452, 2007.

[150] J. Sung, T. Kanade, and D. Kim. A unified gradient-based approach for combining ASM into AAM. *International Journal of Computer Vision*, 75(2):297–309, 2007.

[151] M. Turk and A. Pentland. Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, 3:71–86, 1991.

[152] P. Unsang and A.K. Jain. Face matching and retrieval using soft biometrics. *IEEE Transactions on Information Forensics and Security*, 5(3):406–415, 2010.

[153] L.J.P. van der Maaten, E.O. Postma, and H.J. van der Herik. Dimensionality Reduction : A Comparative Review. *Journal of machine learning research*, 10:1–41, 2009.

[154] B. van Ginneken, M.B. Stegmann, and M. Loog. Segmentation of anatomical structures in chest radiographs using supervised methods: a comparative study on a public database. *Medical Image Analysis*, 10(1):19–40, 2006.

[155] J. van Os and S. Kapur. Schizophrenia. *Lancet*, 374(9690):635–645, 2009.

[156] A.A. Vinkhuyzen, N.L. Pedersen, J. Yang, S.H. Lee, P.K. Magnusson, W.G. Iacono, M. McGue, P.A. Madden, et al. Common SNPs explain some of the variation in the personality dimensions of neuroticism and extraversion. *Translational Psychiatry*, 2:e102, 2012.

[157] P. Viola and M.J. Jones. Robust Real-time Object Detection. *Proc. of IEEE workshop on Statistical and Computational Theories of Vision*, 2001.

[158] P. Viola and M.J. Jones. Robust Real-Time Face Detection. *International Journal of Computer Vision*, 57(2):137–154, 2004.

[159] N. von Cramon-Taubadel, B.C. Frazier, and M.M. Lahr. The problem of assessing landmark error in geometric morphometrics: theory, methods, and modifications. *American journal of physical anthropology*, 134(1):24–35, 2007.

[160] W.Y.S. Wang, B.J. Barratt, D.G. Clayton, and J.A. Todd. Genome-wide association studies: theoretical and practical concerns. *Nature reviews. Genetics*, 6(2):109–18, 2005.

[161] J.D. Watson and F.H. Crick. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature*, 171(4356):737–738, 1953.

[162] S.M. Weinberg, E.A. Jenkins, M.L. Marazita, and B.S. Maher. Minor physical anomalies in schizophrenia: a meta-analysis. *Schizophrenia research*, 89(1-3):72–85, 2007.

[163] K. Wolffhechel. Can you judge a book by its cover? - an investigation into the proposed connection between looks and personality. *M.Sc.Eng. Thesis*, Technical University of Denmark, 2012.

[164] K. Wolffhechel, J. Fagertun, U.P. Jacobsen, W. Majewski, A.S. Hemmingsen, C.L. Larsen, S.K. Lorentzen, and H. Jarmer. Interpretation of appearance: The effect of facial features on first impressions. *submitted to PLoS ONE*, 2014.

[165] R. Xu and D. Wunsch. *Clustering*. John Wiley and Sons, 2009.

[166] J. Yang, B. Benyamin, B.P. McEvoy, S. Gordon, A.K. Henders, D.R. Nyholt, P.A. Madden, A.C. Heath, et al. Common SNPs explain a large proportion of the heritability for human height. *Nature Genetics*, 42(7):565–569, Jul 2010.

[167] L.A. Zadeh. Fuzzy sets. *Information and Control*, 8(3):338–353, 1965.

[168] F.R. Zahir, A. Baross, A.D. Delaney, P. Eydoux, N.D. Fernandes, T. Pugh, M.A. Marra, and J.M. Friedman. A patient with vertebral, cognitive and behavioural abnormalities and a de novo deletion of NRXN1alpha. *Journal of Medical Genetics*, 45(4):239–243, Apr 2008.

[169] Z. Zhang. Iterative Point Matching for Registration of Free-Form Curves and Surfaces. *International Journal of Computer Vision*, 13(2):119–152, 1994.

[170] Z. Zhang. *Camera Calibration*, chapter 2, pages 4–43. Emerging Topics in Computer Vision. Prentice Hall Professional Technical Reference, 2004.

[171] H. Zou, T. Hastie, and R. Tibshirani. Sparse principal component analysis. *Journal of Computational and Graphical Statistics*, 15:265–286, 2006.